STUDIES IN
LANGUAGE
COMPANION
SERIES 94

Language Complexity

Typology, contact, change

Edited by
Matti Miestamo, Kaius Sinnemäki
and Fred Karlsson

JOHN BENJAMINS PUBLISHING COMPANY

Language Complexity

Studies in Language Companion Series (SLCS)

This series has been established as a companion series to the periodical *Studies in Language*.

Editors

Werner Abraham
University of Vienna

Michael Noonan
University of Wisconsin-Milwaukee USA

Editorial Board

Joan Bybee
University of New Mexico

Ulrike Claudi
University of Cologne

Bernard Comrie
Max Planck Institute, Leipzig

William Croft
University of New Mexico

Östen Dahl
University of Stockholm

Gerrit J. Dimmendaal
University of Cologne

Ekkehard König
Free University of Berlin

Christian Lehmann
University of Erfurt

Robert E. Longacre
University of Texas, Arlington

Brian MacWhinney
Carnegie-Mellon University

Marianne Mithun
University of California, Santa Barbara

Edith Moravcsik
University of Wisconsin, Milwaukee

Masayoshi Shibatani
Rice University and Kobe University

Russell S. Tomlin
University of Oregon

John W.M. Verhaar
The Hague

Volume 94

Language Complexity. Typology, contact, change
Edited by Matti Miestamo, Kaius Sinnemäki and Fred Karlsson

Language Complexity

Typology, contact, change

Edited by

Matti Miestamo
Kaius Sinnemäki
Fred Karlsson
University of Helsinki

John Benjamins Publishing Company
Amsterdam / Philadelphia

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Library of Congress Cataloging-in-Publication Data

Language complexity : typology, contact, change / edited by Matti Miestamo, Kaius Sinnemäki and Fred Karlsson.
p. cm. (Studies in Language Companion Series, ISSN 0165-7763 ; v. 94)
Includes bibliographical references and index.
1. Typology (Linguistics). 2. Languages in contact. 3. Linguistic change. 4. Creole dialects. 5. Pidgin languages.
P204 .L25 2008
410--dc22 2007037811
ISBN 978 90 272 3104 8 (Hb; alk. paper)

© 2008 – John Benjamins B.V.
No part of this book may be reproduced in any form, by print, photoprint, microfilm, or any other means, without written permission from the publisher.

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

Table of contents

Introduction: The problem of language complexity vii
Fred Karlsson, Matti Miestamo & Kaius Sinnemäki

Part I
Typology and theory 1

Complexity in linguistic theory, language learning and language change 3
Wouter Kusters

Grammatical complexity in a cross-linguistic perspective 23
Matti Miestamo

Complexity trade-offs between the subsystems of language 43
Gertraud Fenk-Oczlon & August Fenk

Complexity trade-offs in core argument marking 67
Kaius Sinnemäki

Assessing linguistic complexity 89
Patrick Juola

How complex are isolating languages? 109
David Gil

Complexity in isolating languages: Lexical elaboration versus grammatical economy 133
Elizabeth M. Riddle

Grammatical resources and linguistic complexity: Sirionó as a language without NP coordination 153
Östen Dahl

Part II
Contact and change 165

Why does a language undress? Strange cases in Indonesia 167
John McWhorter

Morphological complexity as a parameter of linguistic typology: Hungarian as a contact language 191
Casper de Groot

Language complexity and interlinguistic difficulty 217
Eva Lindström

Complexity in nominal plural allomorphy: A contrastive survey of ten Germanic languages 243
Antje Dammel & Sebastian Kürschner

Part III
Creoles and pidgins 263

The simplicity of creoles in a cross-linguistic perspective 265
Mikael Parkvall

Complexity in numeral systems with an investigation into pidgins and creoles 287
Harald Hammarström

Explaining Kabuverdianu nominal plural formation 305
Angela Bartens & Niclas Sandström

Complexity and simplicity in minimal lexica: The lexicon of Chinook Jargon 321
Päivi Juvonen

Index of languages 341

Index of authors 345

Index of subjects 349

Introduction

The problem of language complexity

Fred Karlsson, Matti Miestamo & Kaius Sinnemäki
University of Helsinki

1. Language complexity

The topic of language complexity certainly is a timely one. During the past few years it has surfaced in many different contexts: as one of the themes in the pioneering agenda of the Santa Fe Institute founded by Murray Gell-Mann; in Perkins' (1992) study of the complexity of deictic systems in the languages of the world; in Nichols' (1992) profound analysis of global morphological and syntactic diversity; in Hawkins' (1994, 2004) work on the processing-related foundations of grammar; in Deutscher's (2000) book discussing the evolution of sentential complements in Akkadian; in Kusters' PhD (2003) dissertation on the ultimately social bases of complexity differences and language change; in the extensive discussion on the complexity or simplicity of Creoles initiated by McWhorter (1998, 2001); in Dahl's (2004) book on the origin and maintenance of language complexity; in Chipere's (2003) investigations concerning individual differences in native language proficiency; in Everett's (2005) fieldwork and interpretations of the low complexity of the Amazonian Pirahã language; in Haspelmath's (2006) proposal that the concept of markedness can be dispensed with in linguistic theory and replaced as an explanatory principle by more fundamental, substantive factors such as phonetic difficulty and pragmatic inferences; in Karlsson's (2007a,b) work on constraints on syntactic embedding complexity; in the 2007 Winter Meeting of the Linguistic Society of America, where a workshop was devoted to the complexity topic; in the "Workshop on Language Complexity as an Evolving Variable" arranged by the Max Planck Institute of Evolutionary Anthropology, Leipzig, in April 2007; in the "Recursion in Human Languages" conference at Illinois State University, Normal, Illinois, in April 2007; and so forth.

Several comprehensive complexity-related research projects are currently under way or have recently been completed. One of them is the project "The Grammatical Complexity of Natural Languages", sponsored by the Academy of Finland for the years 2003–2006 and located at the Department of General Linguistics, University of Helsinki. This project, the Department of General Linguistics, and the Linguistic Association of Finland jointly organized the conference "Approaches to Complexity in Language" in Helsinki on August 24–26, 2005. The current volume contains a thematic selection of the papers read at that conference.

The subtitle of this volume is *Typology, Contact, Change*. These catchwords describe the commonalities of the present papers. Language contact, especially when extensive L2 learning is involved, is a main source of complexity reduction (grammar simplification). By definition, such processes involve language change. But complexity reduction is actually at the heart of many types of language change, especially in morphology and syntax. Contact-induced grammatical change is likely to produce outcomes simpler (in some sense) than the original ones, affecting thus the overall typology of a language. The classical instantiation of this pervasive tendency is the diachronic strive to re-establish the One-Meaning–One-Form principle in "disturbing" situations where it does not obtain, e.g., in the presence of synchronically "superfluous" morphophonological alternations in inflectional paradigms. Such complexities are likely to arise in situations where the influence of L2 contact is weaker – situations often referred to as "normal" linguistic change, cf. Dahl's (2004) concept of maturation.

It is, however, not a trivial question how this kind of complexity related phenomena could be captured and described in a systematic fashion. What does complexity mean in the first place? Can it be objectively measured at all? Is the old hypothesis true that, overall, all languages are equally complex, and that complexity in one grammatical domain tends to be compensated by simplicity in another? These fundamental questions are addressed by many contributors to this volume.

Somewhat surprisingly, Nicholas Rescher's (1998) seminal philosophical analysis of what complexity is has never come to the fore in linguistic discussions of the issue. As a starter for conceptual clarification, we quote Rescher's (1998: 1) definition: "Complexity is first and foremost a matter of the number and variety of an item's constituent elements and of the elaborateness of their interrelational structure, be it organizational or operational." More specifically, Rescher (ibid. 9) breaks up the general notion of complexity into the following "modes of complexity" (our account is adapted to linguistically relevant matters and somewhat simplified):

1. ***Epistemic modes***
A. *Formulaic complexity*
a. *Descriptive complexity*: length of the account that must be given to provide an adequate description of the system at issue.
b. *Generative complexity*: length of the set of instructions that must be given to provide a recipe for producing the system at issue.
c. *Computational complexity*: amount of time and effort involved in resolving a problem.

2. ***Ontological modes***
A. *Compositional complexity*
a. *Constitutional complexity*: number of constituent elements (such as phonemes, inflectional morphemes, derivational morphemes, lexemes).
b. *Taxonomic complexity* (or *heterogeneity*): variety of constituent elements, i.e., number of different *kinds* of components (variety of phoneme types, secondary articulations, parts of speech, tense-mood-aspect categories, phrase types etc.).

B. *Structural complexity*
a. *Organizational complexity*: variety of ways of arranging components in different modes of interrelationship (e.g., variety of premodification or postmodification alternatives in basic constituent types such as noun phrases; variety of distinctive word order patterns).
b. *Hierarchical complexity*: elaborateness of subordination relationships in the modes of inclusion and subsumption (e.g., variety of successive levels of embedding and modification in phrases, clauses, and sentences; variety of intermediate levels in lexical-semantic hierarchies).

3. ***Functional complexity***
A. *Operational complexity*: variety of modes of operation or types of functioning (e.g., variety of situational uses of expressions; variety of styles and speech situations; cost-related differences concerning language production and comprehension such as Hawkins' 2004 efficiency, etc.).
B. *Nomic complexity*: elaborateness and intricacy of the laws governing the phenomenon at issue (e.g., anatomical and neurological constraints on speech production; memory restrictions on sentence production and understanding).

Rescher's taxonomy provides a background for evaluating the coverage of the papers in this volume. Descriptive, constitutional, and taxonomic complexity all figure prominently in many of the contributions, especially under headings such as "length of description", "overspecification", and "absolute complexity". Organizational complexity surfaces in connection with "structural elaboration", in the form of redundant morphophonological alternations creating intraparadigmatic morphological complexity. Operational (i.e., processing-related) complexity is treated by some contributors under notions such as "relative complexity" or "efficiency". Hierarchical complexity, a central theme in formal linguistics (cf. the Chomsky hierarchy of languages, the most well-known linguistic measure of syntactic complexity), is not treated at any great length, reflecting the non-formalist approaches of the authors

2. An overview of the contributions in this volume

Three thematic areas emerged as focal from the ensemble of the papers, and these areas define the three sections into which the book is divided: "Typology and theory", "Contact and change", "Creoles and pidgins". These are naturally overlapping themes, but the placement of each paper in a given section is meant to reflect the central focus of the paper. The first section, "Typology and theory", unites eight papers concerned with typological comparison and with general theoretical issues such as the definition and measurement of complexity.

Continuing the discussion started in his 2003 doctoral dissertation, Wouter Kusters addresses fundamental issues in research on linguistic complexity. He argues for

a relative approach to complexity, defining as complex those linguistic features that cause difficulties for L2 learners. He introduces the "general outsider" as an ideal L2 learner, neutralizing the effects of the mother tongue and cultural background, and proposes a set of concrete criteria for measuring complexity. The theoretical and methodological discussion is illustrated with a case study on the development of verbal inflection in selected Quechua varieties.

Matti Miestamo broaches a number of theoretical and methodological issues relevant for the cross-linguistic study of grammatical complexity. Two basic approaches to complexity are distinguished: the absolute one where complexity is seen as an objective property of the system, and the relative one: complexity as cost/difficulty to language users. The usability of these approaches in typological studies of complexity is evaluated. Miestamo argues that in typological studies of complexity it is better to focus on specific functional domains that are comparable across languages. Some general criteria for measuring complexity are introduced and evaluated, in particular the criteria of Fewer Distinctions and One-Meaning–One-Form. As for the relationship between complexity and cross-linguistic rarity, it is likely that rarity shows at least some correlation with absolute complexity.

Gertraud Fenk-Oczlon and August Fenk approach complexity from a systemic perspective, continuing their earlier work on systemic typology (e.g., Fenk-Oczlon & Fenk 1999). They argue that languages vary in terms of complexity within subsystems but that trade-offs occur between the subsystems (phonology, morphology, syntax, and semantics). Analysing data on monosyllables they show that syllable complexity, number of syllable types, and the number of monosyllables correlate in statistically significant ways. Some aspects of semantic complexity are also discussed as well as the difficulty of talking about the complexity of rigid word order.

Kaius Sinnemäki scrutinizes the old idea that languages trade off complexity in one area with simplicity in another. He tests this hypothesis with a complexity metric based on the functional load of different coding strategies (especially head marking, dependent marking and word order) that interact in the marking of syntactic core arguments. Data from a balanced stratified sample of 50 languages show that the functional use of word order has a statistically significant inverse dependency with the presence of morphological marking, especially with dependent marking. Most other dependencies were far from statistical significance and in fact provide evidence against the trade-off claim, leading to its rejection as a general all-encompassing principle. Overall, languages seem to adhere more strongly to distinctiveness (overt marking of important distinctions) than to economy (minimization of the use of overt markers).

Patrick Juola discusses several proposed definitions of complexity and shows how complexity can be assessed in these frameworks. As empirical testbed for developing a measure of complexity Juola uses translations of the Bible into English and 16 other languages. The variation in size of the uncompressed texts is substantially more than the variation in size after Ziv-Lempel (LZ) compression. This suggests that much of the variance in document size of the Bible is from the character encoding system. Juola argues that LZ, with its focus on the lexicon and long-term storage and retrieval

of words, is a better model of the underlying regularities of language as expressed in corpora than linear complexity. Translations tend to be more complex (longer) than the original text. Furthermore, morphological and syntactic complexity can be measured by distorting the original text. Languages are about equally complex, but they express their complexity differently at different levels.

David Gil's paper treats the complexity of isolating vs. non-isolating languages. He presents an experimental approach to cross-linguistic analysis of complexity by investigating the complexity of compositional semantics in these language types. Compositional semantics is considered the simpler the more associational operations are available in a language. According to the compensation hypothesis, their availability should be roughly equal in the two language types but Gil's method reveals that isolating languages generally allow for more associational operations than non-isolating languages. He further argues that the possible locus of complexity in isolating languages could not reside in pragmatics either, but that they have to be considered overall simpler than non-isolating languages, thus counteracting the claim that all languages are equally complex.

Elizabeth Riddle presents a diametrically opposed view on the complexity of isolating languages. Basing her discussion on data from three isolating languages in southeastern Asia, viz. Hmong, Mandarin Chinese and Thai, she argues that while these languages are indeed simple as regards bound morphology, they show considerably more complexity in lexical categories. Her concept of "lexical elaboration" encompasses phenomena such as classifiers, verb serialization, compounding, and a special type of elaborate expressions found in these languages. The discussion also addresses the fundamental issue of the borderline between lexicon and grammar.

Östen Dahl discusses complexity from the point of view of linguistic resources, that is, the set of possibilities the system offers its users. This aspect was left rather untouched in his (2004) book which laid out a framework for describing the increase of complexity in grammaticalization processes. The author argues that Sirionó, a Tupí-Guaraní language spoken in Bolivia, lacks syntactic NP coordination, wherefore its system offers its users fewer possibilities than many other languages do, and is thus less complex in this respect. He interprets the strategies employed by Sirionó as incremental and those resembling the English *and*-strategy as compositional, arguing for a grammaticalization path and a complexity increase from the former to the latter.

The four papers in the section "Contact and change" place their focus on how language complexity is affected in situations of language contact and what happens to complexity when languages change.

John McWhorter continues the discussion started in his 2001 article in the journal *Linguistic Typology.* He has argued that the grammars of languages may undergo significant overall simplification only through large-scale non-native acquisition by adults. The extreme cases of such simplification are creoles, but less extreme simplification vis-à-vis their closest relatives can be observed in many non-creole languages used as lingua francas (e.g., English, Mandarin Chinese, Persian or Indonesian). In his contribution to the present volume, he discusses a few cases found in Indonesia that seem to present challenges to his views.

Casper de Groot argues that Hungarian spoken outside Hungary manifests an increase in complexity compared to Hungarian spoken in Hungary – typical of languages spoken in multilingual environments (cf. Nichols 1992). He shows that the general trend in the contact varieties is towards analyticity and periphrastic expressions and that the newly acquired structures are usually based on ones already present in the replica languages. De Groot adopts the methodology in Dahl (2004) arguing that these changes in Hungarian spoken outside Hungary manifest an overall decrease in morphological complexity which, however, is accompanied by concomitant changes in syntactic complexity.

Eva Lindström's contribution is based on a total of around two years of fieldwork on the island of New Ireland in Papua-New Guinea. The author distinguishes between complexity as an objective property of the system, and difficulty experienced by adult learners of a language, and introduces explicit criteria for approaching both complexity and difficulty. She compares the Papuan language Kuot (her field language) to neighbouring Austronesian languages, discussing the interesting sociolinguistic situation where Kuot is being viewed as more difficult than its neighbours by the communities, and the use of Kuot is declining more rapidly than that of the neighbouring languages.

Antje Dammel and Sebastian Kürschner present an in-depth survey of morphological complexity of nominal plural marking in ten Germanic languages. They propose a metric that combines qualitative criteria, e.g., number of allomorphs, with more qualitative criteria, such as deviations from One-Meaning–One-Form, motivated e.g., by Natural Morphology. The combination of the multiple criteria shows the greatest accumulation of the most complex values in Faroese whereas the simplest values cluster the most in English. The authors further seek to validate the results by psycholinguistic experiments, a welcome appeal to bridge the so-called absolute and relative approaches to complexity (see Miestamo's and Kusters' papers in this volume).

The final section, "Creoles and pidgins", contains four papers focusing on different issues of complexity and simplification in these contact languages, a central question being whether the contact past of these languages shows in their degree of complexity.

Mikael Parkvall addresses the complexity of creoles vs. non-creoles with the help of the database in the *World Atlas of Language Structures* (Haspelmath et al. (eds) 2005) – the largest typological database publically available to date. A method of quantification is introduced, and after selecting the suitable features, a large sample of languages is placed on a scale of complexity. Additional data on a number of pidgins and creoles is analysed and these languages are also placed on the scale. The author approaches his results from different viewpoints present in current creolistics and discusses their implications to the debate on creole origins. In addition to being the first large-scale attempt to quantify the typological profile of creoles vs. non-creoles, it is also a nice illustration of the unprecedented possibilities the WALS database offers for cross-linguistic research in different domains.

Harald Hammarström's paper deals with the complexity of numeral systems and includes a case study of the domain in pidgins and creoles. This domain may be studied in a very broad cross-linguistic perspective since its semantics are exceptionally clear-cut and data is available from a very large number of languages – the array of languages

considered in the paper is impressive even for a typologist reader. Hammarström examines the complexity of numeral systems, defining complexity in terms of minimal description length. The complexity of pidgin and creole numeral systems is compared to their respective lexifiers and to the estimated cross-linguistic average. The paper is interesting in showing how a specific domain may be studied in terms of complexity, and it also connects to the on-going debate on the complexity of creoles vs. non-creoles.

Angela Bartens and Niclas Sandström apply the morphosyntactic 4-M model developed by Carol Myers-Scotton and Janice Jake to Cape Verdean Creole Portuguese (CVC). The 4-M model is claimed to be a universal model for the classification of morphemes according to their degree of cognitive activation in the process of utterance formation. CVC has been less strongly restructured than so-called prototypical creoles. The main topic is nominal plural marking, where CVC has morphosyntactic configurations similar to Brazilian Vernacular Portuguese (BVP). Previous accounts of CVC and BVP nominal plural marking mention the occurrence of (at least) one inflectional marker per noun phrase. Bartens and Sandström demonstrate that the reduction of inflectional plural marking in CVC results in massive loss of morphosyntactic complexity, due to CVC having arisen through substantial reduction and restructuring during creolization and to being young in comparison to its lexifier Portuguese.

Päivi Juvonen examines how the minimal lexicon of the pidgin language Chinook Jargon (CJ) gains maximal efficiency when used in a contemporary fictional text. CJ is well documented and even though its glory days with some 100,000 speakers in the 1880s are long gone, there are still some speakers left in British Columbia and Northwest Oregon. A speaker of, e.g., Swedish may have a passive vocabulary of some 60,000 words whereas many pidgin languages only have vocabularies comprising a few hundred words. The total number of lexical morphemes listed in different sources for such well documented pidgin languages as Chinook Jargon, Mobilian Jargon and Lingua Franca range from around 500 to just over 2000. The paper describes the CJ lexicon from a structural point of view and then examines the use of multifunctional lexical items in comparison to English. The results show, first, that there is no bound morphology (neither derivational nor inflectional) in the variety studied and, second, that there is much more multifunctionality in the pidgin text than in the English texts.

3. Acknowledgements

First and foremost, we would like to thank the authors for their interesting contributions and for the smooth and pleasant cooperation. Each paper has been read and commented on by at least two referees, one being an author of another paper in this volume and the other(s) external. We wish to express our gratitude to the colleagues who have devoted their time for refereeing the papers: Wayan Arka, Peter Bakker, Walter Bisang, Lauri Carlson, Michael Cysouw, Guy Deutscher, Bernd Heine, Magdolna Kovács, Ulla-Maija Kulonen, Randy LaPolla, Pieter Muysken, Peter Mühlhäusler, Johanna Nichols, Jussi Niemi, Timo Riiho and Geoffrey Sampson. Three reviewers preferred to remain anonymous. We would

also like to thank the editors of the Studies in Language Companion Series, Werner Abraham and Michael Noonan, for inviting us to publish this collection of articles in the series and for all their work behind the scene. Thanks are due to Patricia Leplae, Kees Vaes and the other people at Benjamins who have helped in making this book a reality. We are grateful to Kirsi L. C. Reyes, and Taina Seiro and Mette Sundblad for their help with compiling the indexes. The Linguistic Association of Finland deserves special thanks for cooperation in organizing the conference that gave birth to the volume. We would also like to thank the conference participants for the lively and sophisticated scholarly exchange and the nice atmosphere they managed to create with their presence at the conference. For financial and institutional support, we would like to thank the Academy of Finland and the University of Helsinki.

References

Chipere, N. 2003. *Understanding Complex Sentences. Native Speaker Variation in Syntactic Competence*. Basingstoke/New York: Palgrave Macmillan.

Dahl, Ö. 2004. *The Growth and Maintenance of Linguistic Complexity*. Amsterdam: Benjamins.

Deutscher, G. 2000. *Syntactic Change in Akkadian. The Evolution of Sentential Complementation.* Oxford: Oxford University Press.

Everett, D.L. 2005. Cultural constraints on grammar and cognition in Pirahã. Another look at the design features of human language. *Current Anthropology* 46: 621–646.

Fenk-Oczlon, G. & Fenk, A. 1999. Cognition, quantitative linguistics, and systemic typology. *Linguistic Typology* 3: 151–177.

Haspelmath, M. 2006. Against markedness (and what to replace it with). *Journal of Linguistics* 42: 25–70.

Haspelmath, M., Dryer, M, Gil, D. & Comrie, B. (eds) 2005. *The World Atlas of Language Structures*. Oxford: Oxford University Press.

Hawkins, J.A. 1994. *A Performance Theory of Order and Constituency*. Cambridge: Cambridge University Press.

Hawkins, J.A. 2004. *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

Karlsson, F. 2007a. Constraints on multiple initial embedding of clauses. *International Journal of Corpus Linguistics* 12(1): 107–118.

Karlsson, F. 2007b. Constraints on multiple center-embedding of clauses. *Journal of Linguistics* 43(2): 365–392.

Kusters, W. 2003. Linguistic Complexity: The Influence of Social Change on Verbal Inflection. PhD Dissertation, University of Leiden. Utrecht: LOT.

McWhorter, J. 1998. Identifying the creole prototype: Vindicating a typological class. *Language* 74: 788–818.

McWhorter, J. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology* 5(2/3): 125–166.

Nichols, J. 1992. *Linguistic Diversity in Space and Time*. Chicago: The University of Chicago Press.

Perkins, R.D. 1992. *Deixis, Grammar, and Culture*. Amsterdam: John Benjamins.

Rescher, N. 1998. *Complexity. A Philosophical Overview*. New Brunswick and London: Transaction Publishers.

PART I

Typology and theory

Complexity in linguistic theory, language learning and language change

Wouter Kusters
Meertens Instituut

In this paper I discuss how the notion of complexity can be defined and operationalized to serve as a concept in linguistic research domains like typology, historical linguistics and language contact and acquisition studies. Elaborating on earlier work (Kusters 2003) I argue that a relative notion of complexity is to be preferred over an absolute one. With such a substantial notion, I show that possible objections raised against the concept of complexity are not valid. I work this further out for complexity in verbal inflectional morphology. Finally I demonstrate some intricacies of complexity with examples from variation and change in Quechua varieties.

1. Introduction

In this paper I discuss the notion of complexity and give an example of how it can be used in historical and contact linguistics.

The idea that all languages are complex, though some languages more complex than others, meets with quite some resistance. As in this allusion to Orwell's Animal Farm, many linguists and non-linguists alike consider the property of "being complex" as equal to "having a high value (in a cognitive, social or cultural sense)". Indeed, music with complex harmonies or literary texts with complex interpretations are often appreciated more than their opposites. Such aesthetic considerations supported by chauvinistic reasons led many 19th century linguists to suppose that languages like Latin, Sanskrit and German would be more complex, and that more complex languages would relate to more complex, higher cultures. A reaction to such Eurocentric views came with early American anthropology and linguistics, as exemplified by Sapir's (1921: 219) frequently quoted passage:

> Both simple and complex types of language of an indefinite number of varieties may be found spoken at any desired level of cultural advance. When it comes to linguistic form, Plato walks with the Macedonian swineherd, Confucius with the head-hunting savage of Assam.

For a long period of time, notions of complexity would no longer be used in discussions of the cultural embedding of language.

However, after the expansion and specialization of linguistics into subbranches in the second half of the 20th century, the term complexity is no longer avoided in all contexts. For example, in second language research, Klein and Perdue (1997) wrote on the differences between basic, communicatively barely efficient languages of second language learners and complete languages, as spoken by native speakers and used the term complexity in that context. In dialectology complexity and simplification are also common terms (cf. Andersen 1988; Trudgill 1986), while in contact linguistics (e.g., Mühlhäusler 1974; Winford 1998) simplification is used as a notion without immediate association to evaluative judgements as long as it does not take a too prominent position. However, when discussing languages as wholes, and when considering one language as more or less complex than another, many linguists still think that this idea stems from ill-informed laymen at its best, or is based on nationalistic or eurocentric assumptions at its worst. Nevertheless, in the last years various articles have been published, in which differences of overall complexity between languages are discussed (cf. e.g., Braunmüller 1984, 1990; Dahl 2004; McWhorter in Linguistic Typology 2001, and the comment articles therein; Thurston 1982, 1987, 1992; Trudgill 1992, 1996, 1997, 2001).

In this paper I argue for a feasible definition of complexity, that does justice both (1) to the subbranches of linguistics as described above, where complexity is already used as a term, (2) to laymen who hold intuitive ideas about simple and complex languages, as well as (3) to linguistic theory.

I first spend some words on the reasons why the notion of complexity is, or at least, has been so controversial (Section 2). Next I discuss the difference between "absolute" and "relative" notions of complexity, and argue for the latter (Section 3). In Section 4, I unfold my definition of complexity and comment on some objections against it. In Section 5, I further develop the notion of "complexity" in order to give an example of how it can be applied to variation in Quechua verb morphology in Section 6.

2. Absolute and relative complexity

In the modern literature on linguistic complexity there are two positions. On the one hand complexity is used as a theory-internal concept, or linguistic tool, that refers only indirectly, by way of the theory, to language reality. On the other hand, complexity is defined as an empirical phenomenon, not part of, but to be explained by a theory. This difference, and the confusion it may lead to, can be compared for instance with the difference between a theory that *uses* the concept of emotion to explain – possibly non-emotional – behaviour, and a physiological theory *about* emotions as empirical phenomena, which explains them possibly in non-emotional terms. This distinction should be carefully kept in mind in next description and comparison of the two positions.

The oppositional terms of absolute and relative complexity are coined by Miestamo (2006), who uses the term "absolute complexity" as *not* related to the experiences of a particular kind of language user. Instead, absolute complexity is considered to be an

aspect of a language as an autonomous entity. Since we can discuss such an aspect of an abstracted state of language fruitfully only within some kind of conceptual framework, the notion of absolute complexity turns out to be determined by descriptions and theories of language(s). In this sense complexity and related notions like markedness or naturalness have been widely used in various frameworks (cf. Dahl 2004).

For instance, in level-based phonology we may conjecture that a phonological description that needs an extra level is more complex than one that is describable with only one level. Then we may distinguish complex and simple phonologies in terms of this particular framework, depending on the amount of extra machinery or levels needed. However, whether the actual sounds and structures of the languages in question are also experienced as more complex in terms of acquisition speed, production ease or any other measure remains a separate question, as also Miestamo concedes (2006: 4).

So far so good, however, the distinction between the absolute and relative approach is less clear-cut than it seems. Consider another example: in generative grammar it is assumed that children acquire grammars by filling in parameters. Now, it has been suggested (cf. Hyams 1987) that each parameter may have a default, or unmarked value, and a more marked value, that may be called more complex. Some claim that SVO order is less marked than any other order (cf. Gadelii 1997), because SVO order would be the result of an unmarked, default, or *not complex* parameter setting. What particular parameter this would be in this case is not at stake here. In this example, again, a language as autonomous entity may be characterized as having a complex or simple grammar in terms of the theory about the language. Again, whether the production of SVO orders would be faster or otherwise easier is irrelevant.

In this last example, however, we can no longer keep away from the problem which involves the *status of a grammar*. When a grammar is considered to be only a tool of a linguist to provide an appealing, short and clear description of a language, then indeed the complexity of the grammar has nothing to do with possible complexities in language use. This position with respect to language and grammar is rather awkward, however. Since the 1960s most linguists agree – if not in practice then at least in principle – that grammars model knowledge architecture, cognitive make-up or processing features of language users. In generative grammar it has been stated most explicitly that the quest for universals is *not* a quest for statistic generalizations in the void, but a search for the real innate language acquisition device. Language universals are not only universals in the mind of the linguist, but also found in language users' minds. Although less explicitly, in other frameworks too it is common knowledge that what is described is some aspect of language users.

If so, all parts of our linguistic theories should refer – in the end – to aspects of language users. Still remaining within the framework of generative grammar and the example above, we must assume that the difference between an "unmarked" parameter setting in SVO against a marked parameter setting in e.g., OVS, is reflected in the child's acquisition process, although this may be in a quite indirect way. Nevertheless, when it is claimed that a linguistic theory models aspects of the cognitive or, more

specifically, linguistic make-up of a speaker's mind, the notions within the theory should be reflected in the data as well.

Proponents of the absolutist approach do not necessarily use the term "complexity" in this way. When discussing language features from a typological perspective the relation between linguistic (i.e., typological) theory and actual language use may indeed be indirect or apparently irrelevant. Regularities and patterns may be found on a highly abstract level. For instance, SVO languages are more likely to also have the order Noun-Genitive than Genitive-Noun. It is not immediately clear how such tendencies would be properties of the individual language user. It may even be claimed that these are "emergent properties" that can not be reduced to the domain they originate from. When discussing language on such an abstract level it may be tempting to suggest that theoretical (typological) concepts are independent from actual language use or speakers' minds. If so, complexity may refer, for instance, to the number of rules that is needed for a particular description. We could call this approach *formal absolutist*, and when using it consequently it is a tenable position. For instance, in terms of mathematical theory some grammars can be called more complex, in the sense of needing more complex rules. A context-sensitive grammar is in this sense more complex than a context-free grammar, and a rule that says: "yes/no questions are related to their affirmative counterpart by a complete reversal of all syllables" would be simpler than a rule that relates these two by assuming complex syntactic structures. By defining complexity in this manner we have a term at our disposal by which we can handle and compare various formal theories. However, one cannot have it both ways. When defined in a *formal absolutist* way, it is only a sheer coincidence if complex grammars are also experienced by language users as complex. The formal absolutist simple rule for making yes/no-questions does not exist in human language, and may be considered to be too complex in a more substantial (relativist) way. Defenders of the absolutist formal approach should have no problems with that fact.

If so, the meta-level of the description is clearly separated from the object level. With the term complexity, however, it is felt as somehow a draw-back of the theory, if complexity is *only* a theory-internal notion. This can be illustrated with an example of two theories explaining the same phenomenon. In Quechua varieties there is variation in the inflectional verbal conjugation (see also Section 6). Quechua varieties vary both in the categories expressed, as well as in the way they are expressed. Now, the verbal inflection of Bolivian Quechua – genetically closely related to Cuzco Quechua – has been discussed in two morphological theories. Van de Kerke (1996) discusses Bolivian Quechua verbs within the theory of Distributed Morphology, while Lakämper and Wunderlich (1998) discuss it within their constraint-based minimalist morphology. Leaving out the numerous intricacies of Quechua inflection and the theories involved, the following examples are relevant:

Bolivian Quechua:

(1) *maylla-wa-rqa-yku*
wash-1-PAST-1.PL
'You(sing/plur)/he/they washed us.'

Cuzco Quechua:

(2) *maylla-wa-rqa-nki-ku*
wash-1-PAST-2-PL
'You (sing/plur) washed us.'

(3) *maylla-wa-rqa-n-ku*
wash-1-PAST-3-PL
'He/they washed me/us.'

Now we have two theories according to which we can decide what morphology is simple. According to a theory that focuses on the way syntactic features and morphological spell-out are related, the first example would be more complex. This is argued by Van de Kerke (1996), when he discusses the Bolivian data in the framework of Distributed Morphology. The problem with the Bolivian Quechua data would be, then, that the feature of [+ 1st person] is expressed twice, both in the *wa*-morpheme, as well as in the pluralizer *yku*. From the perspective of Distributed Morphology the Cuzco data are more straightforwardly explained. Subject and object affixes have their own slots, and are spelled out once. Van de Kerke (1996: 130) says:

> We have seen that Cuzco Quechua has a very transparent Agr/Tense system. . . it complies with the Mirror Principle in realising a good match between the order of morphemes and the morpho-syntactic categories expressed. However, this ideal transparency has become opaque by a minor reinterpretation of first and third plural marking in Bolivian Quechua, which not only led to a great number of underspecified and doubly specified surface realizations, but even to the realization of subject markers as pluralizers in the case of *-yku* as in *-wa-yku* [unspecified sub- > 1Plob].

Other morphological theories are less focused on syntactic-morphological (mis) matches, and more on the lexicon-morphology-interface. Lakämper and Wunderlich (1998) use such a "pre-syntactic lexicalist theory". These theories load up the lexical material with features that are dealt with in syntax in other approaches. When morphologies are evaluated in such theories, problems – or, "complexities" – as described above count less. Instead, the number of allomorphs, homonyms and fused morphemes become relevant. From that perspective Lakämper and Wunderlich (1998: 147) draw the following conclusion about Quechua complexity:

> . . . Only when we come to the post-Cuzco stages (like Potosí [this is Bolivian Quechua, WK] or Santiago del Estero) . . . has this kind of deficiency been overcome and an affix-oriented system is produced. However, there can only be small changes, and the new system has to work with the affix material inherited from the former stages. As we have seen, the potentially symmetric system that emerges in the most recent stages of Quechua is not ideal either.

Lakämper and Wunderlich acknowledge that Bolivian Quechua is not ideally simple. However, they assess the complexities of Bolivian Quechua in a quite different way. The double spell-out of 1st person in 1) is not a problem for them, as it is for Van de Kerke, while the number of affixes is more of an indicator to them.

The point of this discussion is that on the basis of the same data it is possible to arrive at different conclusions about what forms are as complex. And in that sense, "absolute" complexity is not so absolute as the name suggests. Absolute complexity, in fact, depends on the theory and perspective from which we look at a particular language or language phenomenon.[1]

The absolutist may either accept the theoretical diversity, and select a theory. Then complexity remains firmly part of a particular linguistic theory, claims no link with empirical complexities in itself, and becomes a formal absolutist notion. Otherwise he may try to make an informed choice in what theory describes the data best. That is, he may confront the two claims on what counts as complexity, with what complexity is *in the empirical data*. In the latter case, the absolutist becomes in fact a relativist, since he makes his theoretical description and notion of complexity depend on some kind of *real* complexity.

3. Definition of complexity

3.1 Points of relativity

I will now turn to the relativist position and how it can be further elaborated. For the relativist, the main question is, relative to whom or what? The relativist position always assumes a perspective or a point of view from where a certain phenomenon is evaluated as complex or simple. Actually, the absolutist position can be considered as a relativist position in disguise. The absolutist position may be characterized as relativist since complexity is evaluated from the point of view of a particular linguistic theory. As discussed above, complications in the absolutist position arise when we acknowledge that a linguistic theory is a model of a language user.

What are the possibilities for a relativist definition? We may choose various kinds of language users and various aspects of language. We could examine what kinds of sounds in Russian are complex for adult Chinese classroom learners. We could examine what morphology of Latin is most complex to handle for readers of Cicero. Or we might consider what kinds of word orders are difficult for speakers of SVO languages when confronted with a VSO language.

---

1. Special cases are languages about which most or perhaps even all linguistic theories come to the same conclusions with respect to complexity. As noted by a reviewer, Ndyuka will be estimated as less complex than Kabardian under any linguistic analysis. Proponents of complexity theory may adduce such extreme examples to defend their idea. However, the displacement of the notion of complexity from any theoretical framework leaves them defenseless on the broad middle-ground of languages, where the measure of complexity is dependent on more fine-grained theoretical decisions and the perspectives of the language users chosen, which leads necessarily to a notion of "relative" instead of "absolute" complexity.

My conception of complexity elaborates on the naive language user's intuitions when he wonders about what languages are complex or simple. This naive language user conceives other languages as non-native languages, and he has an outsider perspective on the speech community of the language in question. Then we still have numerous conceptions of complexity: a native speaker of Chinese will count other phenomena as complex in German, Vietnamese or Swahili, in comparison with a speaker of Turkish or Warlpiri. In order to construct a general notion of complexity, I have defined an ideal language user type, namely a "generalized outsider". This person speaks a first language, and is not familiar with the second language in question, nor with the customs and background knowledge of the speech community. He or she is primarily interested in using language for communicative purposes in the restricted sense of the word, and not in all kinds of verbal play, ritual exchanges, and poetic uses. We model this person as a *generalized* outsider, in order to prevent either facilitating or hampering influences of the first language on acquiring the second language.[2]

Now, I define complexity as the amount of effort a generalized outsider has to make to become acquainted with the language in question. I distinguish three dimensions of language processing where outsider complexity plays a role.

A generalized outsider learns the language in question at a later age, and is not a native speaker. Therefore, phenomena that are relatively difficult for a second language learner in comparison with a first language learner are the most complex. Phenomena that are easy to acquire for a second language learner but difficult for a first language learner are the least complex. The notion "generalized" prevents positive and negative interferences of an accidental first language to cloud our account of complexity.

A generalized outsider does not have much shared (linguistic and non-linguistic) background knowledge with other members of the speech community. She will therefore be helped with a language that is relatively easy to perceive and understand. In contrast, production difficulties of a language are less hindering for an outsider, because she may adapt the difficult production phenomena to some kind of imperfect second language form, while she can of course not force other native speakers to adapt. A language in which perceptual processing is relatively easy in comparison with production is therefore less complex under my definition.

---

2. As several commentators remarked, the notion of a *generalized outsider* may be just as problematic an abstraction as Chomsky's ideal speaker-listener in a homogeneous speech community. However, this notion seems to me to be useful; perhaps it is not valid as an abstraction or generalization over actual concrete cases of *outsiders*, but it is still useful as a kind of most extreme type, or possibility. Indeed, full objectivity as a "view from nowhere" is impossible in science (cf. Nagel 1986), and a complete outsiders' perspective is even more improbable in actual speech communities. Nevertheless, such concepts may play at least an heuristic role in theory building.

A generalized outsider is primarily interested in clear transmission of information, and less interested in learning a language for all kinds of symbolic meanings, like expressing personal and group identity and aesthetic feelings. Phenomena that are not or less functional in this respect are therefore more complex under my definition.

When pondering on what parts of language to select for a more detailed examination, several considerations play a role (cf. Kusters, 2003: 12–21), and I have chosen the domain of verbal inflectional morphology. From a practical point of view complexity in inflection is easily assessed cross-linguistically. In most descriptive grammars information about verbal inflection is given. Secondly, it is more independent from pragmatics, discourse, register, or inter-speaker variation than other linguistic domains, and is rather uniform across various situations. Moreover, it occurs in most sentences, and must be learned and used by all kinds of language users. Nevertheless, it varies considerably *between* languages: some languages have extensive inflectional systems, while others have no inflection at all.

In comparison with derivational morphology, inflectional morphology is less involved with questions of lexical semantics or indeterminacy with respect to the predictability of meaning. In addition, inflectional morphology concerns morphological structure, which is, with its idiosyncratic allomorphies, exceptions, and communicatively superfluous distinctions, less susceptible to trade-off effects than syntax or phonology. Inflection is also quite easy to define independently from theoretical considerations. Between morphological theories there is rather much agreement at least about what counts as inflection. Finally, verbal inflection stands out above other kinds of inflection, since it is most wide-spread both within and across languages.

3.2 Objections against outsider complexity

Now that we have a more detailed definition of complexity, we can answer some objections that have been raised against the notion that there could be differences in complexity.

Many of the arguments are a priori, and boil down to the claim that all languages are necessarily equally complex, because complexities in one domain of language are balanced by simplicity in another domain. For instance, Aitchison (1991: 214) writes:

> A language which is simple and regular in one respect is likely to be complex and confusing in others. There seems to be a trading relationship between the different parts of the grammar which we do not fully understand.

What is usually lacking is a specification of what complexity means. When it is "outsider complexity", most of the objections against the relevance of the concept evaporate, as I will show, while other objections were already incoherent by themselves. Since these arguments are met again and again, lets spend some words on following these arguments to their (absurd) conclusions.

Focusing on Aitchison's trading relationship lets try to defend it by supposing that it is due to limitations on mental capacities; the mind/ brain can process and contain only a limited amount of complexity. An increase in complexity in one component, then, necessarily leads to a decrease elsewhere. This argument presupposes that the average speaker already uses the maximal amount of the "complexity space". Accepting this for the sake of the argument, other problems remain. When assuming that mono-linguals use the complete "complexity space", the acquisition of foreign languages becomes difficult to explain. It would entail either that the acquisition of a second language has the effect that the native language becomes less complex, or that the acquisition of a second language occupies a quite different "complexity space".

Still playing the devil's advocate, and assuming that for each language there is a separate and restricted mental space, problems remain. That is, when we conceive language competence as consisting of lexical and grammatical knowledge, it is hard to imagine how lexical knowledge would be always equally complex. For example, knowing approximately 66.000 words, as Shakespeare had allegedly at his disposal (cf. Efron and Thisted 1976), would in some way have to be equally complex as knowledge of a very rudimentary lexicon restricted to, say, small-village pre-modern agrarian life.[3] One might be inclined to argue that indeed these two kinds of knowledge are equally complex, because with both kinds of lexicons it is possible to *function* in certain niches in society. However, in that case we reach the bottom-line of argumentation, since this entails that all languages that exist must be equally complex just because they exist, and therefore "function". According to my definition of complexity however, a large lexicon, ceteris paribus, is probably more complex than a smaller lexicon, since an outsider has less problems in acquiring and using a smaller lexicon, cf. however, the discussion by Fenk-Oczlon & Fenk (this volume), Juvonen (this volume) and Riddle (this volume).

Another argument is that the trade-off lies in language use somehow: when in one domain of a language complexity disappears, somewhere else new complexities would have to surface. Languages would always remain on the same level of complexity. That is, a simple language cannot exist, because different language components have counteracting "preferences" for simplicity. A structure that is easily produced would be complex to perceive and vice versa. The force of such arguments depends partly on the definition of complexity. For complexity, as I define it, perceptual simplicity outweighs articulatory smoothness. Therefore, when the trade-off is argued to lie between different kinds of language use/users, the objection does not hold, since in my definition complexity refers to only one kind of language user: the outsider.

---

3. It might be argued that pre-modern farmers do not have necessarily small lexicons. However, that would be missing the point: an advocate of equi-complexity should give no empirical evidence of equally large lexicons, but must give a priori arguments why differences in lexicon size are impossible.

A valid argument against the concept of complexity needs to show that there is *always* a trade-off between different domains. However, this seems not to be the case. For instance, the replacement of a vast array of lexically conditioned plural markers as in Classical Arabic by a system with transparent optional plural marking as in KiNubi Arabic does not seem to involve any trade-off. Whether Classical Arabic plural marking is indeed more complex (for an outsider) than KiNubi Arabic plural marking must be shown in empirical tests, but this possibility cannot be excluded beforehand. With respect to a trade-off between phonological components, there are indications (cf. Maddieson 1984: 23) that even the opposite may be the case.

Another flaw in a priori arguments based on a mysterious trade-off is that they say nothing about complications, only about simplifications. It is less plausible that a complication in one part, for instance, the borrowing of words with previously unknown phonemes, necessarily results in a simplification elsewhere. When the trade-off would only apply to simplifications, differences in complexity cannot be excluded by the trade-off argument.

It may be maintained that there must be a trade-off *somewhere* in the language, even if we do not know where. The simplification of plural marking in the example above, may lead to more difficulties in the interpretation of plural forms. The argument in its general form runs as follows: in some languages complexities may be found on a tangible linguistic level. In other languages complexities emerge in an undefined part of the "pragmatic component". In such arguments the trade-off function extends from core language competence to wider communicative and cultural competence.

Such equi-complexity seems not very likely, as long as it is not shown how mediations in complexity between dispersed linguistic domains could take place. When the trade-off would be supposed to lie between a limited number of semantically or functionally related domains, the claim would be more reasonable (cf. Sinnemäki's article in this volume). Although the argument of *a priori* equi-complexity cannot be excluded, it is not falsifiable: for every change in complexity it can be argued that there is another component, in some hitherto unknown domain of language structure, pragmatics, or culture, where the amount of complexity would be leveled out.

4. Methodology

4.1 Three aspects and three dimensions

Now we have discussed and defined complexity, how can we further operationalize it in order to study it in the wild?

First of all we must examine what aspects of inflectional morphology are in fact experienced as troublesome by this "generalized outsider". For that purpose I have distinguished three aspects of the language outsider: the outsider as a second language learner, as a hearer, and as someone mainly interested in the communicative instead

of the more symbolic aspects of language use. I also distinguish three aspects in inflectional morphology (cf. Kusters 2003):

1. The number of categories expressible in the verb; I call this *Economy*. The less inflectional categories and category combinations a language allows, the more economical the language is. For instance, Ecuadorian Quechua is more economical than Cuzco Quechua, since it does not express 2nd person object agreement nor 1st person plural object agreement.
2. The transparency of the expressions of the categories, which I call *Transparency*. The more deviations there are from a most transparent structure, in terms of allomorphy, fusion, fission (that is, the opposite of fusion: one meaning, expressed in two morphemes), and homonymy, the less Transparency. For instance, Argentinean Quechua is more transparent than Bolivian Quechua since it expresses 2nd person object agreement more often with a uniform affix, instead of a fused affix.
3. The consistency within the order of the expressed elements; I name this *Isomorphy*. The more the order in the morphological domain is computable and motivated by the order in the semantic or syntactic domain, the more isomorphic the morphology is. Ecuadorian Quechua is more Isomorphic than Cuzco Quechua, since it has a univocal affix order, that mirrors a semantic-syntactic order elsewhere, while the description of Cuzco Quechua morpheme order needs some extra stipulations.

When we want to find out what outsider complexity is with respect to inflectional morphology, we must examine empirical data on the difficulties of each of the three outsider aspects (second language learning, perception, communicative use) crossed with each of the three morphological dimensions, which yields nine classes. Here the absolute and relative approaches fundamentally differ. In the relative approach, it is in the end an empirical question whether an "outsider" has more problems with Ecuadorian Quechua, Bolivian Quechua, Turkish or Russian inflection. Nevertheless, in practice many of the hypotheses of the absolutist approach about the difficulties of irregularities correspond to empirical findings. This is, however, not so by definition. In the relative approach we can – and in fact we do – distinguish between the problems of first language learners, and compare these with the problems of second language learners. The relative approach allows us to say that, for instance, little Economy is, quite unexpectedly, found to be easy for L1 learners, while it is difficult for L2 learners and "outsiders" (cf. Kusters, 2003: 48ff.).

Empirical investigations into the difficulties of the nine sorts of language processing, are unfortunately not as abundant as we would like. There is, for instance, hardly any research in the perceptional difficulties of an isomorphic uniform order in comparison with variant morpheme ordering. When we try to gather and use the empirical data in these domains as much as possible we come to the following conclusions about what counts as difficult. For the research and data on which this table is based I must refer to my dissertation (Kusters 2003: 45–63).

Table 1. Schematized representation of preferences for inflectional phenomena in various processing dimensions (+ = preference, 0 = neutral and -, --, and --- = degrees of difficulty)

	Speaker	**L1 learner**	**Symbolic use**	**Hearer**	**L2 learner**
Redundant agreement	--	+	0	+	--
Non-redundant agreement	--	+	+	-	--
Aspect/Tense/Mood	-	+	+	0	-
Voice	0	+	+	+	-
Morphological allomorphy	-	--	+	-	---
Accidental homonymy	0	--	0	--	---
Fission	-	-	+	+	--
Fusion	+	0	0	0	-
Phonological allomorphy	+	-	+	0	--
Structural homonymy	+	+	0	0	+
Isomorphy	+	0	0	+	+
Marked affix order	0	0	0	0	0
Inconsistent affix order	-	0	0	-	-

4.2 From the individual to the social

The next step is to examine what kind of languages are simpler or more complex. A straightforward hypothesis is that a language (as a social fact, that is, as "E-language") becomes simpler the more it has been confronted with demands of generalized outsiders. Assuming that the actual state of a language (E-language) is the product of all language users, who produce it on the basis of how they got acquainted with the language in question, it is not too far-fetched to assume that a history of a speech community with many of these "generalized outsiders", effects in the end that the language adapts to the outsiders' preferences. For instance, we expect that the history of intense language contact between coastal urban Norwegian in the late Middle Ages and North-Germanic dialects has had a *leveling,* or *pidginising*, or *koineising*, that is, *simplifying* effect on Norwegian.

As we measure complexity relative to the generalized outsider, we expect that in the end the measure of complexity of full languages (E-languages), correlates with the amount of outsiders in the speech community history. To speak more smoothly about the speech communities, I propose that there are two extremes of speech communities, Type 1 and Type 2, of which the latter is an extreme "generalized outsider community". We may further characterize these communities as follows:[4]

The population of an idealized Type 1 community is relatively small, and most people know each other. Most interactions take place among members of the community,

4. Note that the various aspects of the two types (e.g., size, kind of network structure, amount of contact with outsiders) are not independent but mutually reinforcing. This is a disadvantage from an analytical point of view. It is however an advantage with respect to its description from

and in interactions with outsiders a different language is used. Members of a Type 1 community share a large common background. Life cycles are relatively predictable, originality and innovation are not appreciated very much, and neither is the transmission of new information. Language is used to keep social relations between members of the community in balance. There is a body of literature (written or oral) in the community with a sacrosanct status. People are proud of their language, and they have stories that relate the origin and form of the language to their religion and their cultural origins. In everyday life, verbal play and language skills are appreciated. Different registers exist but there is no dialectal variation. Ties between community members and the possession of communal values are strong and no local centers of prestige develop, and therefore, also no dialects.

Type 2 communities only form a speech community because all members use the same language. They do not necessarily form a unit in space or time, or share social or cultural values. The number of speakers of a Type 2 language can be high. Most speakers know other languages as well, and the language in question is often not their first language, but only used as a lingua franca. The members of such communities do not share much background knowledge, and their precise way of speaking the language may differ. In interactions the language is mostly used for negotiating and exchanging practical information. The language is not associated with a cultural or religious standard, and speakers accommodate their way of speech freely in order to be clearly understood. The language is not a medium in which identities are expressed. There are no registers, but there may be different dialects, or different ways of speaking, possibly related to the original languages of the speakers.

In order to examine complexity differences most fruitfully it is easiest to compare related languages, and related speech communities, which have a common mother language, and only a different social history. If so, we may conjecture that when a language splits, and one variety becomes more like a Type 1, and the other like a Type 2 community, we expect that the latter becomes simpler in its inflectional morphology.

5. Case study: Bolivian, Argentinean and Ecuadorean Quechua

5.1 Social history of three Quechua varieties

I will now show how this turns out for one case of inflectional changes in related varieties: the Quechua languages of Peru, Argentina and Ecuador. Instead of expounding a wealth of tiny examples from diverse language families, I concentrate on

an interpretive point of view. Compare also the traits of a "personality type" in classical psychology, which are not mutually independent, though mutually reinforcing the saliency of the type.

one particular case. This provides the opportunity to search for the precise linguistic locations of complexity; to follow the exact paths of simplification among related varieties, and to find out how the actual implementation of the theory fares. In addition, by focusing on Quechua, we will find out that simplification can mean something quite different from the Indo-European based notion of "erosion", that commonly occurs at word boundaries. I will first describe some features of the histories of the Quechua varieties, and next compare their morphologies (for a more comprehensive treatment, cf. Kusters, 2003: 249–303 and references therein).

The origin of Quechua lies on the coast of central Peru before 500AD. Some varieties, called Quechua II, spread slowly southward along the coast, while Quechua I moved over the highlands. There was a second split before 1500, when traders moved northwards along the coast as far as Ecuador and used Quechua as a lingua franca (which became known as Quechua IIb or Ecuadorian Quechua). Meanwhile Quechua IIc was spoken near Ayacucho and Cuzco, which would grow in power as the capital of the Incas. Quechua spread to Bolivia and Argentina under the Incas, while it was learned in special schools, in ethnic interactions and in official Inca religious and administrative transactions. After the Spanish Conquest in the 1530s the status of Quechua changed. Initially the Spanish government promoted Quechua, but later in the 18th century Quechua was associated with backwardness and its existence came under threat.

In the first decades after the Conquest, Quechua was used as the lingua franca in southern Peru by Andean migrants for communicative purposes and as a church language. Later Spanish became more important. Today, in spite of centuries of repression of the Andean culture and extensive migrations and depopulations, Quechua is still widely spoken. Cuzco Quechua has become the prestige language – both spoken and written – for an indigenous upper class of Andean nobles and it is supported by a language academy. We expect some early simplification in Cuzco and Ayacucho varieties, because of the turbulent history of migration and population changes. In comparison with the other Quechua varieties, we expect less inflectional change, since in Peru it was not adopted as a second language or lingua franca by large groups of L2 learners.

After arrival in Argentina under the Incas Quechua further spread and was consolidated under the Spaniards. The heterogeneity of speakers who took Quechua there was much higher than in Peru. First and second language speakers of Quechua, Andean migrant workers from diverse Quechua varieties and other Andean languages, Spanish colonists and Spanish priests were all involved in its spread and development. Argentinean Quechua became a mix of different native, koineized and second language varieties of Quechua. It does not have a written literary tradition, and has far less status than Spanish. Because of the quite rapid L2 acquisition process, the wide diversity of speaker backgrounds involved in its early spread, and its continuous low prestige, we assess the Argentinean Quechua speech community as far closer to a Type 2 community than the one in Peru.

In pre-Inca times Quechua was adopted in Ecuador as a trade language, while the second Quechua wave came through the Inca expansion, by which Inca Quechua

spread over the first layer of Quechua IIb speakers, and gained new Andean L2 speakers from various backgrounds. After an initial decrease due to the decline in population during and after the Inca civil war Quechua began to extend its domain as the indigenous lingua franca that expressed Andean identity. Among the Quechua communities discussed here, the Ecuadorian one displays most traits of a Type 2 speech community because of its early history as a trade language.

5.2 Quechua inflection

The total number of categories in a word is high in Quechua, while the number expressed in each affix is low. The inflectional categories on which I focus are tense, subject agreement, object agreement and number. The general order in verbal morphology in all Quechua II varieties is: VERB – OBJ -TENSE – SUB – NUM. For example, in Cuzco Quechua:

(4) *Yanapa-wa-rqa-nki-cis.*
help-1-PAST-2-PL(EXCL)
'You (plur) have helped me.'

Quechua varieties deviate in different ways from this template. Some categories appear in fused forms, sometimes the order deviates, and sometimes the meaning of the affixes is dependent on the meaning of other affixes.

In the next tables, I show the past tense paradigms of Cuzco, Ayacucho (in italics when different from Cuzco), Argentinean and Ecuadorian Quechua. Each column refers to a different object, while each row refers to a different subject. The endings which these different subject-object combinations display are spelled out in the cells.

Table 2. Cuzco and Ayacucho Quechua past tense inflection

	1SG	2SG	1PL.INCL	1PL.EXCL	2PL	3/NO OBJ
1 SG	*	rqa-yki	*	*	rqa-yki-cis	rqa-ni
2 SG	wa-rqa-nki	*	*	wa-rqa-nki-ku	*	rqa-nki
3 SG	wa-rqa-n	rqa-sunki *su-rqa-nki*	wa-rqa-ncis	wa-rqa-n-ku	rqa-sunki-cis *su-rqa-nki-cik*	rqa
1PL.INCL	*	*	*	*	*	rqa-ncis
1PL.EXCL	*	rqa-yki-ku	*	*	rqa-yki-ku	rqa-y-ku *rqa-ni-ku*
2PL	wa-rqa-nki-cis	*	*	wa-rqa-nki-ku	*	rqa-nki-cis
3PL	wa-rqa-n-ku *wa-rqa-n*	rqa-sunki-ku *su-rqa-nki*	wa-rqa-ncis	wa-rqa-n-ku	rqa-sunki-cis *su-rqa-nki-cik*	rqa-ku

Cuzco and Ayacucho Quechua differ with respect to the possibilities of the plural 3rd person marker, the 1st person plural marker, and the position of the *su*-affix (i.e., its fusion with the *nki*-affix).

Table 3. Argentinean Quechua past tense inflection (differences with Peruvian Quechua are shaded).

	1SG	2SG	1PL.INCL	1PL.EXCL	2PL	3/NO OBJ
1SG	*	su-ra-ni/ ra-yki	*	*	ra-yki-cis	ra-ni
2SG	a-ra-nki	*	*	a-ra-yku	*	ra-nki
3SG	a-ra	su-ra	a-ra-ncis	a-ra-yku	su-ra-nki-cis	ra
1PL.INCL	*	*	*	*	*	ra-ncis
1PL.EXCL	*	su-ra-yku	*	*	ra-yki-cis	ra-yku
2PL	a-ra-nki-cis	*	*	a-ra-yku	*	ra-nki-cis
3PL	a-ra-nku	su-ra-nku	a-ra-ncis	a-ra-yku	su-ra-nki-cis	ra-nku

Apart from phonological differences, Argentinean differs with respect to: (1) the reanalysis of *su*, this affix has turned into a 2nd person object marker, (2) the reanalysis of *y-ku* and *n-ku* into fused morphemes, which prevents separate *ku* from appearing. As a result of these minimal changes, maximal paradigmatic effects occur (cf. Kusters 2003: 276ff.). Most notable are the less transparent ways to express 1st person plural exclusive objects, and the more transparent way to express a 2nd person object.

As discussed in Section 3, it is not completely straightforward how to assess the changes in such paradigms, but taking everything into account our impression is that this paradigm is a little simpler than the Peruvian varieties.

Table 4. Ecuadorian Quechua past tense inflection

	1SG	2SG	1PL.INCL	1PL.EXCL	2PL	3/NO OBJ
1SG	*	*	*	*	*	-rka-ni
2SG	-wa-rka-ngi	*	*	*	*	-rka-ngi
3SG	-wa-rka	*	*	*	*	-rka
1PL.INCL	*	*	*	*	*	-rka-nci
1PL.EXCL	*	*	*	*	*	-rka-nci
2PL	-wa-rka-ngi-ci	*	*	*	*	-rka-ngi-ci
3PL	-wa-rka-(kuna)	*	*	*	*	-rka-(kuna)

This paradigm is univocally simpler than the Peruvian ones, on all three morphological dimensions of Section 5.1. Affixes that do not express a clear unambiguous meaning (*yki, sunki*) are lost, and in addition, the plural marker has a more local scope: it can only pluralize its adjacent affix.

5.3 Conclusions and interpretations

We expected Cuzco Quechua to be more like a Type 1 language (that is, language characteristic for a Type 1 community) than Argentinean Quechua, and Ecuadorian

Quechua to be most like a Type 2 language. We also expected that all Quechua varieties would display Type 2 language characteristics because of the turbulent Andean history both under the Incas and after the Spanish Conquest.

The complex morphology of Cuzco and Ayacucho Quechua has, however, not essentially changed. How to explain this? This conservatism may be a reaction of the Quechua speakers to outside pressure exerted on their culture and language. Perhaps the new learners of Quechua from other Indian communities could not exert their influence on Quechua, because of the prestige of Quechua or because of the possibly relatively low numbers of learners. Furthermore, perhaps the stability on household level and intra-generational language transmission remained fully intact. However, perhaps the clear difference in the extent of simplification between e.g., Scandinavian and Quechua can be explained by the agglutinative structure of Quechua verbs, which may be intrinsically more stable than Old Norse structure (cf. Kusters 2003 for a far more extensive discussion).

Ecuadorian Quechua simplification can be explained by the following factors: (1) Unlike other varieties Ecuadorian Quechua was heavily influenced by traders who used Quechua as a lingua franca. Quechua was then acquired and used by adults who were not corrected by a strong Quechua norm. (2) The L2 learners of Quechua in Ecuador had probably more diverse and more deviant language structures than those in Argentina and certainly in Peru. In other cases of simplification as in Swahili, L1 similarity also determines the extent of simplification. Here we see the relevance of various senses of "relative complexity". A "generalized outsider" may have problems with Quechua that the actual Quechua learners in Peru did not have, because of their L1 background.

In Argentina, *su* is on its way to becoming a transparent object marker. This extension and regularization of object agreement may be an autonomous change, but because of its exceptionality in Quechua varieties, it counts as a simplifying change. As far as we may speak then of Argentinean simplification, it is interpreted in terms of the diversity of language backgrounds in Argentina, and the little contact Argentina had with the Quechua heartland.

However, Argentinean Quechua is less Transparent and Economical than Ecuadorian Quechua in its verbal inflection, although both countries seem to display similar social historical patterns. A plausible explanation is that all speakers of Ecuadorian Quechua were initially second language learners, while in Argentina there was a core of native speakers, who could function as a norm for "correct" Quechua. Argentinean Quechua served as a means of inter-ethnic communication and was not used exclusively for trading purposes or only by foreign language learners.

From studies in Arabic and Scandinavian we know that simplification may also be analysed as incidental effects of language-internal changes, like stress shift and concomitant vowel reduction and loss of affix distinctions. Such language-internal explanations are implausible for Quechua. The Quechua affixes vanished quite abruptly, and there were no intermediate stages in which there was a phonologically reduced form. Moreover, the affixes that disappeared in Quechua were not word-final, and they disappeared without

any other stress shift. Cuzco Quechua even contains a shift in metrical structure which is similar to that of Germanic, but still did not erode at the end of the word.

6. Conclusions

In spite of the controversies of the notion, this article – as well as the other articles in this volume – shows that "complexity" is a viable notion in linguistics. When properly defined, there are no a priori arguments why we could not profit from this concept. The issue whether complexity is best used as an absolute or a relative notion, is complex itself. I have argued that when we base our idea of complexity on the experiences of a particular type of language user, we receive more insights from subfields like theoretical linguistics, creole studies, contact linguistics, historical linguistics and typology.

When we apply the concept of complexity to an actual case of language change – in this article, change in Quechua verbal inflection – we may conclude that languages are adapted in their level of complexity to the preferences of language users. We also find that these preferences do not immediately translate in simplifications but are mediated by the characteristics of the structure of the language in question. The case of Quechua is interesting in this respect, because it shows what kinds of simplification are possible besides the well-known cases of Indo-European language change, which are all too often plainly characterized as "erosion". While "erosion" brings notions of natural (i.e., causal) change with it, the kind of simplification in Quechua invokes ideas of teleology and goal-directedness in matters of language change.

As usual in models that take social and historical facts into account, predictions do not come out as clear-cut as we might wish. Broad generalizations have all kinds of stubborn contours at closer look. Nevertheless, when we further examine the relevant social-historical factors and the morphological structures and confront these with empirical data, we may arrive at a new typology of simplification paths in the future.

Abbreviations

1	first person
2	second person
3	third person
EXCL	exclusive
INCL	inclusive
OBJ	object
PL	plural
PST	past
SG	singular

References

Aitchison, J. 1991. *Language Change: Progress or Decay?* Cambridge: Cambridge University Press.

Andersen, H. 1988. Center and periphery: Adoption, diffusion and spread. In *Historical Dialectology: Regional and Social*, J. Fisiak (ed.), 39–83. Berlin: Mouton de Gruyter.

Braunmüller, K. 1984. Morphologische Undurchsichtigkeit – ein Charakteristicum kleiner Sprachen. *Kopenhagener Beiträge zur germanistischen Linguistik* 22: 48–68.

Braunmüller, K. 1990. Komplexe flexionssysteme – (k)ein Problem für die Natürlichkeitstheorie? *Zeitschrift für die phonetische Sprachwissenschaft und Kommunikationsforschung* 43: 625–635.

Dahl, Ö. 2004. *The Growth and Maintenance of Linguistic Complexity.* Amsterdam: John Benjamins.

Efron, B. & Thisted, R. 1976. Estimating the number of unknown species: How many words did Shakespeare know? *Biometrika*, 63 (3): 435–437.

Gadelii, K.E. 1997. *Lesser Antillean French Creole and Universal Grammar* [Gothenburg Monographs in Linguistics 15]. Gothenburg: Gothenburg University.

Hyams, N. 1987. The Theory of Parameter and Syntactic Development. In *Parameter Setting*, T. Roeper & E. Williams (eds), 91–121. Dordrecht: Reidel.

Kerke, S.C. van de 1996. Affix order and interpretation in Bolivian Quechua. PhD Dissertation, University of Amsterdam.

Klein, W. & Perdue, C. 1997. The basic variety (or: Couldn't natural languages be much simpler?). *Second Language Research* 13: 301–347.

Kusters, W. 2003. Linguistic Complexity; the Influence of Social Change on Verbal Inflection. PhD Dissertation, University of Leiden.

Lakämper, R. & Wunderlich, D. 1998. Person marking in Quechua: A constraint-based minimalist analysis. *Lingua* 105: 113–148.

Maddieson, I. 1984. *Patterns of Sound.* Cambridge: Cambridge University Press.

McWhorter, J. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology* 5: 125–166.

Miestamo, M. 2006. On the feasibility of complexity metrics. In *FinEst Linguistics, Proceedings of the Annual Finnish and Estonian Conference of Linguistics, Tallinn, May 6–7, 2004* [Publications of the Department of Estonian of Tallinn University 8], K. Kerge and M.-M. Sepper (eds), 11–26. Tallinn: TLÜ.

Mühlhäusler, P. 1974. *Pidginization and Simplification of Language.* Canberra: Linguistic Circle.

Nagel, T. 1986. *The View from Nowhere.* Oxford: Oxford University Press.

Sapir, E. 1921. *Language.* London: Harcourt, Brace & World, Inc.

Thurston, W.R. 1982. *A Comparative Study of Anem and Lusi.* Toronto: University of Toronto.

Thurston, W.R. 1987. *Processes of Change in the Languages of North-Western New Britain* [Pacific Linguistics B 99]. Canberra: Australian National University.

Thurston, W.R. 1992. Sociolinguistic typology and other factors effecting change in north-western New Britain, Papua New Guinea. In *Culture Change, Language Change; Case Studies from Melanesia* [Pacific Linguistics C 120], T. Dutton (ed.), 123–139. Canberra: Australian National University.

Trudgill, P. 1986. *Dialects in Contact.* Oxford: Blackwell.

Trudgill, P. 1992. Dialect typology and social structure. In *Language Contact, Theoretical and Empirical Studies*, E.H. Jahr (ed.), 195–211. Berlin: Mouton de Gruyter.

Trudgill, P. 1996. Dialect typology: Isolation, social network and phonological structure. In *Festschrift for William Labov*, G. Guy (ed.), 3–22. Amsterdam: John Benjamins.

Trudgill, P. 1997. Typology and sociolinguistics: Linguistic structure, social structure and explanatory comparative dialectology. *Folia Linguistica* 31: 349–360.

Trudgill, P. 2001. Contact and simplification: Historical baggage and directionality in linguistic change. *Linguistic Typology* 5: 371–374.

Winford, D. 1998. Creoles in the Context of Contact Linguistics. Paper presented at the symposium Pidgin and creole linguistics in the 21st century, New York, January 1998.

Grammatical complexity in a cross-linguistic perspective

Matti Miestamo
University of Helsinki

In this paper, I address theoretical and methodological issues in the cross-linguistic study of grammatical complexity. I identify two different approaches to complexity: the absolute one – complexity as an objective property of the system, and the relative one – complexity as cost/difficulty to language users. I discuss the usability of these approaches in typological studies of complexity. I then address some general problems concerning the comparison of languages in terms of overall complexity, and argue that in typological studies of complexity it is better to focus on specific domains that are comparable across languages. Next, I discuss a few general criteria for measuring complexity. Finally, I address the relationship between complexity and cross-linguistic rarity.

1. Introduction

The study of language complexity has recently attracted a lot of interest in the field of linguistic typology. The consensus that all languages are equally complex is being challenged by a growing number of authors (e.g., McWhorter 2001; Kusters 2003; Shosted 2006). This paper addresses theoretical and methodological issues concerning the study of grammatical complexity, especially from a cross-linguistic perspective.[1] Complexity has been approached either from the absolute or the relative point of view in linguistics; Section 2 will introduce these perspectives and discuss their pros and cons in typological research. In Section 3, I will address two general problems concerning the comparison of languages in terms of global (overall) complexity, viz. the problems of representativity and comparability. Especially the latter problem leads to the conclusion that typological studies of grammatical complexity should focus on specific domains that are cross-linguistically comparable. Functional domains provide feasible tertia comparationis for such studies. A few general criteria of complexity introduced in the literature are discussed in Section 4, and two general principles for

1. I wish to thank Guy Deutscher, Fred Karlsson, Masja Koptjevskaja-Tamm, and Kaius Sinnemäki for valuable comments on the manuscript. The support of the Academy of Finland (under contract 201601) and the Universities of Helsinki and Antwerp is gratefully acknowledged.

approaching the complexity of functional domains are proposed. Section 5 addresses the relationship between complexity and cross-linguistic rarity, discussing it in the light of absolute vs. relative complexity. The most important points made in the paper are summarized in Section 6.

2. Absolute complexity and (relative) cost/difficulty

In Miestamo (2006a,b), I identified two different points of view from which complexity has been approached in linguistics: the absolute (theory-oriented, objective) approach defines complexity in terms of the number of parts in a system, whereas the relative (user-oriented, subjective) one defines it in terms of cost and difficulty to language users. In this section I will discuss the two approaches at more length, paying special attention to their applicability to cross-linguistic studies of complexity. Note that Dahl (2004: 25–26) uses the terms absolute and relative complexity in a different sense, but these terms do not play a very prominent role in his approach to complexity.

An absolute approach is adopted, e.g., by McWhorter (2001, 2007, this volume) and Dahl (2004). The basic idea behind the absolute approach is that the more parts a system has, the more complex it is. To give a simple example, a language that has 34 phonemes, e.g., Kwazá (Kwaza; van der Voort 2004: 45–46), has a more complex phoneme inventory than one that only has 18, e.g., Tauya (Trans-New Guinea, Madang; MacDonald 1990: 21–31).[2] Obviously, the idea behind counting parts of systems does not always mean looking at lists of elements that make up an inventory, and the parts are not always as straightforwardly countable – this idea is to be taken in a more general sense. The same notion is behind, e.g., seeing a high number of interactions between the components of a system as increasing complexity (cf. Fenk-Oczlon & Fenk, this volume).

Information theory (beginning with Shannon 1948) provides some basic principles that allow us to make the idea of "number of parts" more generally applicable. In accordance with the basic principles of the complexity measure known as Kolmogorov complexity (see Li & Vitányi 1997), Dahl (2004: 21–24) argues that the complexity of a linguistic phenomenon may be measured in terms of the length of the description of that phenomenon; the longer a description a phenomenon requires, the more complex it is – the 34-member phonemic inventory of Kwazá requires a longer description than the Tauya system with 18 phonemes. A less complex phenomenon can be compressed to a shorter description without losing information. On a high level of abstraction we

2. These are the numbers of indigenous (non-borrowed) phonemes in these two languages according to the authors of the grammars. The names and genealogical affiliations of the languages mentioned in the paper follow the classification by Dryer (2005).

may say that we are still dealing with the number of parts in a system, but these parts are now the elements that constitute the description of the system. Applications of the notion of Kolmogorov complexity often make use of computerized compression algorithms; Juola (1998, this volume) provides an example within linguistics (see discussion below). I will adopt the basic idea on a very general level, and will argue for its usability with descriptive tools developed by linguists rather than mathematicians or computer scientists. Furthermore, since we focus on grammatical complexity here, we will be concerned with what Gell-Mann (1994) calls effective complexity; this notion pays attention to the length of the description of the regularities or patterns that an entity, e.g., the language system, contains, leaving everything that shows no regularity or patterning outside its scope.[3]

A relative approach is adopted, e.g., by Kusters (2003, this volume) and Hawkins (2004). This approach defines complexity in terms of cost and difficulty to language users, i.e., how difficult a phenomenon is to process (encode/decode) or learn. The more costly or difficult a linguistic phenomenon is, the more complex it is according to the relative view. A central issue to be taken into account in this approach is that a phenomenon that causes difficulty to one group of language users (e.g., hearers) may facilitate the task of another group (e.g., speakers). In his study of the complexity of verbal inflection, Kusters (2003: 51–52, 56–57; this volume) examines several phenomena that occur in inflection and discusses them from the point of view of different types of language users. He argues, based on different psycholinguistic studies of processing and acquisition, that, for example, redundant agreement is difficult for speakers and L2 learners, but facilitates the task of hearers and L1 learners, while fission (one meaning expressed by many forms syntagmatically) aids hearers but causes difficulties to the three other types of language users.

Given this, the question "complex to whom?" is central to the relative approach to complexity. Whether a phenomenon is to be seen as simple or complex, depends on whether one takes the point of view of the speaker, hearer, L1 acquirer or L2 learner. The approach that one chooses to take naturally depends on the goals that one's research has. Kusters (2003: 6–7, this volume) defines complexity in relation to adult learners of language, and therefore those properties of languages that cause difficulties for L2 learners are defined as complex. The simplifying effects of sociolinguistic factors such as language contacts are a central topic in Kusters' study, and the difficulties experienced by adult learners of language are therefore important; the primary relevance of L2 learners is clear in this case. However, the primacy of L2 learners is by no means obvious if we are

3. Grammar is the domain of regularities and patterns, whereas a description of the lexicon has to resort, to a much larger extent, to merely listing the elements that the system contains. In that sense, the characterization of the complexity of the lexicon, is more about plain description length ("pure" Kolmogorov complexity), and the notion of effective complexity plays a much less important role.

looking for a maximally general view of language complexity. They could in fact be considered the least important of the four groups mentioned – every natural language has (or has at least had) speakers, hearers, and L1 learners, but L2 learners are in most cases a much more marginal group of language users. The latter may form the majority of the users of a given language as is the case in English and Swahili, but this is an uncommon situation in the world's languages, and even in these cases it would be hard to argue for the position that this group of language users is in some sense primary in comparison to native speakers. If we were to choose a group of language users whose difficulties general statements of the complexity of a given language should be based on, L2 learners would certainly not be the first user type to think of.

The question "complex to whom?" causes a general problem for a relative approach to complexity. There will always be some conflict between definitions of complexity based on difficulty for different groups of language users. No general user-type-neutral definition of complexity is possible. One of these groups can be chosen as criterial for the definition of complexity in studies such as Kusters (2003), where the primacy of adult learners of language has a clear motivation; similarly, if we are using the concept of complexity for the purposes of a study of L1 acquisition, we can use a definition of complexity relative to L1 learners. But for a general approach to complexity, such a choice cannot be made. Furthermore, both L1 and L2 learners are also speakers and hearers, and even if we choose either group of learners as criterial, situations will arise where speaker and hearer perspectives will be in conflict. To define relative complexity in more general terms, all groups of language users should be taken into account. There should be a way of measuring the difficulty of each linguistic phenomenon under study for each group of users and then seeing what the contribution of each type of difficulty would be to the general complexity of the phenomenon. Such a measure is clearly not possible. It is true that for some phenomena there will be less conflict between the different groups of users – e.g., structural homonymy is either neutral to or preferred by all the user types mentioned by Kusters (this volume, Table 1) – but in general, the relative approach to complexity cannot avoid this problem.

There is a second general problem that a relative approach to complexity has to face. Estimations of the cost or difficulty of each piece of language for the different groups of language users can only be made on the basis of psycholinguistic studies of processing and acquisition. However, the availability of such research is far from sufficient for this purpose. Kusters (this volume) also acknowledges the problem, saying that there is not enough research that could be used to characterize the difficulty that all the aspects of verbal inflection that he studies cause for different types of language users. When we move beyond verbal inflection, the magnitude of the problem naturally increases. In general, when looking at language complexity from a broad cross-linguistic point of view, we are likely to encounter a large number of phenomena for which we simply cannot determine with satisfactory accuracy what kinds of difficulties they cause to different groups of language users.

Given these problems, I suggest that in general, and especially in broad cross-linguistic studies, complexity be defined in absolute terms. The absolute approach to

complexity allows us to leave these problems aside and define complexity in more objective terms – objective in the sense that it does not depend on any particular group of language users or on incomplete knowledge of cognitive processes. Complexity should therefore be defined, to put it in the most general terms, as the number of parts in a system or the length of its description. Whether and how the different aspects of complexity defined this way contribute to difficulty for different types of language users is a very important question, but a separate one, to be addressed with the help of psycholinguists studying language processing and acquisition.

Dahl (2004: 39–40) suggests that the term complexity be entirely reserved for an objective (information-theoretic, i.e., absolute) conception of complexity, and when talking about cost and difficulty, one should use the terms cost and difficulty, not complexity. This is indeed a terminologically sound approach – especially since complexity (in the absolute sense) is a widely used concept in theories of complexity and information in many different fields of scholarship – and will reduce misunderstandings about the nature of complexity. In complexity theories that have applications across disciplines, complexity is defined objectively as a property of systems, and interesting observations can then be made as to what consequences the increase or decrease of complexity in a system has, or what the causes behind the increase or decrease of complexity are. Keeping the concepts of complexity and cost/difficulty apart in linguistics, enables us to make such observations, e.g., between grammar and psycho- or sociolinguistic phenomena, benefiting our understanding of both linguistic systems and psychological and societal phenomena. Absolute complexity connects to what complexity means in other disciplines and thus also opens possibilities for interdisciplinary research.

As I suggested above, in some cases it may be possible and theoretically motivated to define complexity in relative terms, as in Kusters' study where the effect of L2 learners to simplification is central; however, using the terms cost and difficulty instead of complexity when cost and difficulty are meant and reserving the term complexity for absolute complexity would not take anything away from such studies. The relationship between (absolute) complexity and cost/difficulty is an important question, and approaching this question will become easier and more straightforward if the notions are kept apart by the use of clear terminology. McWhorter has a similar sociolinguistic agenda as Kusters, arguing for contact-induced simplification, but he chooses to define complexity in clear absolute terms. For Lindström (this volume), both complexity and difficulty are important but the concepts are explicitly kept separate. In the remainder of this paper, I will use complexity in the absolute sense, unless otherwise noted, and the terms cost and difficulty instead of relative complexity in appropriate places.

Description length depends on the linguistic theory in terms of which the description is made. Theories differ, and the phenomena to be described also differ in the sense that for the description of some phenomena there will be much more agreement among linguists than for the description of some others – linguists will probably find it easier to agree on the number of aspectual categories in a language than on the description of

a syntactic phenomenon like passivization.[4] As we will see in Section 4 below, interesting cross-linguistic studies of complexity may be done using widely accepted theoretical concepts (cf. Dixon's 1997 notion of Basic Linguistic Theory, BLT), and many interesting things can be said about the complexity of different phenomena based on simple and straightforward criteria that are likely to be widely accepted by linguists of different theoretical persuasions. Some linguistic theories aim at psychological reality, e.g., generative grammar when it claims to represent innate principles of human language, or theories of markedness that see a connection between cognitive/conceptual difficulty and linguistic markedness (see Haspelmath 2006 for an overview of the different uses of the concept of markedness). With such theories, description length could, at least in principle, be of some relevance in studying cost and difficulty as well; in practice this naturally depends on how well a given theory can live up to its claims of psychological reality (cf. also Section 5 below). Note also that the use of description length as a complexity measure is not the same thing as Chomsky's (1957: 49–60) evaluation procedure, which compares (the simplicity of) different grammatical descriptions of one and the same linguistic phenomenon (a specific structure or a whole language), whereas the point here is to compare the length of the description of different linguistic phenomena using the same theoretical principles in the description (whichever theory is chosen).

Juola's (1998, this volume) computational approach attempts to quantify complexity in maximally objective terms. He sees complexity in terms of compressability of texts – the shorter the compressed version of a text, the less complex the original. Cross-linguistic comparison is made using parallel corpora with translations of texts in different languages. Compression algorithms (Juola uses zip) operate on repetition of strings of characters, and some aspects of the complexity of word forms can be captured by this approach. Morphological complexity is thus easier to investigate in these terms than the complexity of other domains of grammar, e.g., syntax or the meanings of the grammatical categories expressed. To take an example, the compression algorithms can detect word order patterns only when there are multiple instances of the same lexemes (in the same form) occurring together in similar or different orders, which is clearly not common enough in natural texts. More generally, we may also ask what the application of a mathematical algorithm on linguistic products (texts) can reveal about the complexities of the underlying systems (grammar, lexicon) that are needed to produce these texts.

In this section I have given several arguments for why complexity should be approached from the absolute point of view in cross-linguistic studies. Although the notions of complexity and difficulty are logically independent, and must be kept apart for theoretical and methodological reasons, their relationship is worth examining – complexity does not entail difficulty and difficulty does not entail complexity, but to which

4. One might want to say that absolute complexity is *relative to* the theory chosen (cf. Kusters, this volume), but I wish to reserve this formulation for user-relativity.

extent they correlate in language is an interesting question. In many cases, more complex structures can be expected to be more difficult as well (at least to some groups of language users), which also shows in that similar criteria have been used for measuring complexity and cost/difficulty (cf. Section 4); in each case, however, psycholinguistic studies are needed to verify this. After having examined linguistic phenomena from the point of view of absolute complexity, we may, in many cases, be able to explain our findings in functional terms making use of the notions of cost and difficulty as well.

3. Global vs. local complexity

The question of global (overall) complexity of languages interests both linguists and non-linguists. With the latter, this usually means characterizing entire languages as easy or difficult to learn. Among linguists, the received view is that all languages are, overall, equally complex in their grammars; complexity differentials can be found in different areas of grammar, but complexity in one area is assumed to be compensated by simplicity in another. Hockett (1958: 180–181) formulates this as follows:

> Objective measurement is difficult, but impressionistically it would seem that the total grammatical complexity of any language, counting both morphology and syntax, is about the same as that of any other. This is not surprising, since all languages have about equally complex jobs to do, and what is not done morphologically has to be done syntactically. Fox, with a more complex morphology than English, thus ought to have a somewhat simpler syntax; and this is the case.
>
> Thus one scale for the comparison of the grammatical systems of different languages is that of average degree of morphological complexity – carrying with it an inverse implication as to degree of syntactical complexity.

A more recent formulation can be found in Crystal (1997: 6). As is evident from most of the contributions to this volume, the received view is not shared by all linguists. In recent years, it has been most notably challenged by McWhorter (2001, 2007, this volume) and Kusters (2003, this volume).

McWhorter (2007, this volume) proposes a metric for measuring and comparing the global complexity of languages, in order to study complexity differences between languages restructured and simplified by contact (e.g., creoles) and languages that have not been affected by contact in such a way. The metric pays attention to overt signalling of distinctions beyond communicative necessity on different levels of language. The following three criteria are used (McWhorter, this volume):

1. Overspecification
2. Structural elaboration
3. Irregularity.

Overspecification refers to the "marking of semantic categories left to context in many or most languages, such as evidential marking", and accordingly, designates language A as more complex than language B to the extent that it makes more (unnecessary) semantic/pragmatic distinctions in its grammar. Structural elaboration is about the "number of rules mediating underlying forms and surface forms, such as morphophonemics", the global complexity of a language increasing with the number of rules in its grammar. According to the third criterion, Irregularity, the more irregularities a grammar contains, the more complex it is overall. Note that in an earlier version of the metric (McWhorter 2001), a slightly different set of criteria is used, but the basic idea behind the two versions of the metric is the same. I will come back to the criteria in Section 4.

In Miestamo (2006a), I introduced two general problems that all attempts to measure global complexity will have to face: representativity and comparability. The problem of representativity means that no metric can pay attention to all aspects of grammar that are relevant for measuring global complexity. Even if this were theoretically possible, it would be beyond the capacities of the mortal linguist to exhaustively count all grammatical details of the languages studied, especially in a large-scale cross-linguistic study. This problem is acknowledged by McWhorter (2001: 134). The problem of representativity may perhaps be solved in the sense that it may be possible to arrive at a level of representativity that enables one to identify very clear complexity differences.

The problem of comparability is about the difficulty of comparing different aspects of grammar in a meaningful way, and especially about the impossibility of quantifying their contributions to global complexity. McWhorter's three criteria are not commensurable: imagine language A with a lot of overspecification, language B with extensive structural elaboration and language C with large amounts of irregularity – how do we decide which criterion weighs more? The same problem applies within the scope of each criterion: to take a random example from Overspecification, how to make the number of distinctions made in deictic systems commensurable with that made in tense systems? The problem of comparability concerns the comparison of all systems and subsystems of language regardless of the criteria used in the complexity metric. How should we compare, e.g., syntactic and morphological complexity and quantify their contributions to global complexity?[5]

Comparison of languages in terms of global complexity becomes possible only when the complexity differences are very clear so that the criteria one uses do not give conflicting results. If one has a set of criteria to measure and compare the complexity of languages A and B, and each criterion designates language A as more complex, then

5. In principle, if linguists could one day achieve a comprehensive and completely objective theory of language ("the Theory of Language"), the lengths of the descriptions of different languages in terms of this theory could be used for measuring the global complexity of these languages.

there is no need to assess the contribution of each criterion. McWhorter's metric is designed for bringing out clear complexity differences between languages simplified by contact and languages without a heavy contact past. It seems to work in the cases examined, providing thus a useful tool for the intended purpose. The same can be said about Parkvall's study (this volume), where a group of languages is designated as less complex than other languages by all (or at least an overwhelming majority of the) criteria used. But when the complexity differences of the languages compared is not so clear and the criteria give conflicting results, there should be a way of quantifying the contribution that each criterion makes to global complexity. In the absence of such quantifiability, studies of global complexity can only bring out very clear global complexity differences between languages. Similarily, as to the problem of representativity, the attainable level of representativity may be enough to show differences between the global complexity of the languages when the differences are very clear – as seems to be the case in McWhorter's (2001) comparison of creoles and non-creoles.

As I have argued in Miestamo (2006a), given the problem of comparability, the cross-linguistic study of grammatical complexity should primarily focus on specific areas of grammar, i.e., on local complexity. We can compare only what is comparable. Many areas of grammar can be meaningfully compared across languages and the choice of these areas naturally depends on one's theoretical goals and orientations. The areas to be compared may be essentially formal, such as phonological inventories or morphological systems, or essentially functional, i.e., various functional domains such as (the encoding of) tense, aspect or referentiality. As typologists argue (e.g., Stassen 1985, Croft 2003), cross-linguistic comparability is best achieved in functional terms, and this also provides a good basis for the cross-linguistic study of grammatical complexity. We can thus compare the encoding of similar functional domains across languages and then make cross-linguistic generalizations about the complexities of these domains, provided of course that languages grammaticalize similar domains – some languages grammaticalize tense, others aspect and yet others both; a language where tense is not grammaticalized will naturally not be included in a comparison of the domain of tense. It should be noted that even when focusing on comparable functional domains, we still need to pay attention to the problem of comparability between the different criteria we use to examine the complexity of these specific domains.

Typologists often talk about poor and rich systems, e.g., of tense, but often these are only loose characterizations of the complexity of the domains. More exact quantification of minimal, average and maximal complexity of domains would certainly contribute to a better understanding of these domains in general. It would also allow for the examination of typological correlations between the complexity of different domains, which would provide answers to the question whether complexity in one domain is compensated by simplicity in another. Although testing the equi-complexity hypothesis will hardly be possible at the level of global complexity (cf. above), examining possible trade-offs between specific domains is worthwhile in many respects – not only because it will give us important insights into the relationships between the domains,

but also in the more general sense that it may provide evidence for or against general cognitive mechanisms responsible for such trade-offs (cf. also Sinnemäki, this volume, who examines trade-offs between the use of different coding means within the domain of core argument marking).

4. Criteria for complexity

In Section 2, I discussed the general principles behind the different approaches to complexity. I will now take up some more concrete criteria in terms of which complexity can be and has been approached. McWhorter (2001, 2007, this volume) and Kusters (2003, this volume) have provided the most explicit criteria for comparing grammatical systems in terms of complexity and I will start by discussing their proposals (on a rather general level – for more concrete examples, see their papers in this volume). I will then propose two general principles for measuring complexity from a cross-linguistic viewpoint.

McWhorter (2001: 134–135) explicitly states that his approach to complexity is independent of processing concerns, i.e., he approaches complexity from an absolute point of view. Indeed, if we look at the three criteria used by McWhorter (2007, this volume), briefly introduced in Section 3 above, we can see that they are quite straightforwardly understandable in terms of description length: Overspecification refers to the number of grammaticalized distinctions made and the more distinctions a language makes within a domain – e.g., in the domain of evidentiality, to use the example given by McWhorter – the longer the description of this domain becomes in the grammar of the language. Structural Elaboration is about the number of rules and is directly interpretable as description length. Finally, Irregularity also increases the length of a description: the more irregularity, the longer the list of separate items in the grammar.

The relative basis of Kusters' (2003, this volume) definition of complexity was discussed in Section 2. I will now take a closer look at the actual criteria he uses. The complexity of verbal inflection is measured with the following principles: 1. Economy – restriction of the number of overtly signalled categories, 2. Transparency – clarity of the relation between meaning and form, and 3. Isomorphy – identity of the order of elements in different domains. The principle of Economy is violated when verbal inflection overtly signals agreement or categories like tense, aspect or mood. The principle of Transparency is essentially about the principle of One-Meaning–One-Form, and it is violated by phenomena like allomorphy (one meaning – many forms paradigmatically), homonymy (many meanings – one form paradigmatically), fusion (many meanings – one form syntagmatically) and fission (one meaning – many forms syntagmatically). Finally, the principle of Isomorphy is violated when the order of inflectional affixes expressing given verbal categories is different from cross-linguistic preferences in the mutual ordering of affixes expressing these categories; cross-linguistic preferences are taken to reflect functional-level preferences. Kusters (2003: 45–62) discusses how the

various possible violations of these principles affect different types of language users. I will not go into these details here, but the point is that those violations that cause difficulty to L2 learners (to the extent that psycholinguistic studies are available to assess this) are interpreted as complexity. Note that although the principles are developed for (and used in) a study of the complexity of verbal inflection, it is clear that they can be used for characterizing any kind of morphological complexity; furthermore, as they are quite general in nature, there is no reason why their applicability should be restricted to morphological complexity. I will therefore interpret them as more general criteria of complexity and compare them with those proposed by McWhorter.

It is notable that although McWhorter approaches complexity from the absolute point of view and Kusters from the relative one, the criteria they use are in many respects similar. McWhorter's Overspecification is close to Kusters' Economy. Those violations of Kusters' Transparency that have any regularity would be subsumed under McWhorter's Structural elaboration – phenomena like fusion and allomorphy increase the number of rules mediating underlying forms and surface forms in McWhorter's view. And finally, McWhorter's Irregularity would take care of those violations of Transparency that have no regularity. As these examples show, the concrete criteria used in absolute and relative definitions of complexity may in many cases look very much alike, but as these definitions have different bases, the motivations to use a given criterion and the way it is to be interpreted are different.[6]

In Miestamo (2006b), I proposed that two very general principles can be used as criteria in an absolute approach to complexity, especially when taking functional domains as the point of departure: the principle of Fewer Distinctions and the well-established principle of One-Meaning–One-Form.[7] Violations of these principles increase the complexity of a linguistic entity. The two principles overlap with McWhorter's and Kusters' criteria in many ways. In the following, I will discuss the advantages of the two principles I propose.

When we investigate the complexity of a functional domain in our languages of study, we may pay attention to two aspects of grammar: grammatical meaning and the encoding of grammatical meaning. In investigating grammatical meaning we pay attention to the meanings languages express grammatically within a functional

6. The similarity is understandable since absolute and relative approaches to complexity work within the same theoretical traditions – similar theoretical concepts are interpreted against different backgrounds. Relative approaches have not founded linguistic theory anew, based purely on psycholinguistic experiments, and furthermore, the experiments referred to are themselves naturally also made within existing linguistic theories.

7. The One-Meaning–One-Form principle, known under this name since Anttila (1972), is sometimes referred to as the principle of isomorphy in the literature; to avoid confusion with Kusters' principle of Isomorphy, I will not follow this usage here.

domain irrespective of the formal means by which these grammatical meanings are encoded. In the study of the encoding of grammatical meaning we pay attention to the relationship between the meanings and the forms that encode them. Methodologically, the investigation of grammatical meaning precedes the investigation of the encoding of grammatical meaning in a cross-linguistic study of the complexity of a functional domain – we must first investigate what meanings are expressed before we can pay attention to the formal details of their expression. The principles of Fewer Distinctions and One-Meaning–One-Form are applicable to the study of grammatical meaning and the encoding of grammatical meaning, respectively. It should perhaps be emphasized that the principles are by no means restricted to morphology but can be used for examining morphosyntactic phenomena in general.

The principle of Fewer Distinctions can be seen in terms of description length as follows: language A, where more functional distinctions are grammaticalized within a given functional domain, requires a longer description for that functional domain than language B, where fewer distinctions are made; language A thus shows more complexity in this respect. To take a simple example, Hdi (Afro-Asiatic, Biu-Mandara) with its three tense categories – (referential) past and two futures (Frajzyngier 2002) – is less complex in this respect than Nasioi (East Bougainville) with its elaborate metric tense system – two futures, present, and three past tenses, to mention only the categories that are not combinations of tense and aspect in Hurd and Hurd (1970). The length of a description depends on the theory in terms of which a phenomenon is described, but the more a theory allows languages to be described in their own terms and not in terms of categories imposed by other languages, the more probable it will be that a large number of tense distinctions, for example, will require a longer description than a small one. (The interaction of the tense system with other functional domains may of course be more complex in the language with a smaller number of distinctions.)

The One-Meaning–One-Form principle can be connected to complexity in the absolute sense, since a situation where the morphosyntactic coding of a function strictly adheres to the One-Meaning–One-Form principle can be given a shorter description than one where the principle is violated. When the form-function correspondences are not one-to-one, either syntagmatically or paradigmatically, the description of the system needs additional specification concerning these form-function relationships. I will now illustrate this with examples from the expression of negation in declarative clauses (references to the original sources are given with the examples, but for analysis, see also Miestamo 2005).

Syntagmatic violations of the One-Meaning–One-Form principle can be found in Kiowa (1) and Kemant (2).

(1) Kiowa (Kiowa-Tanoan) (Watkins 1984: 158, 204, 214)
a. *k'yą́·hį̂·* *Ø-cán*
man 3SG-arrive.PFV
'The man came.'

b. *hɔ́n mátʰɔ̀n Ø-cą́ˑn-ɔ̂ˑ kʰíˑdêl-gɔ̀ˑ*
NEG girl 3SG-arrive.PFV-NEG yesterday-since
'The girl hasn't come since yesterday.'

(2) Kemant (Afro-Asiatic, Central Cushitic) (Appleyard 1975: 333–334)

a. *was-* 'hear', IMPF

	AFF	NEG
1SG	*wasäkʷ*	*wasägir*
2SG	*wasyäkʷ*	*wasäkar*
3SG.M	*wasäkʷ*	wasäga
3SG.F	*wasät(i)*	*wasäkäy*
1PL	*wasnäkʷ*	*wasägənir*
2PL	*wasyäkʷən*	*wasäkänar*
3PL	*wasäkʷən*	*wasägäw*

b. *was-* 'hear', PFV

AFF	NEG
wasəɣʷ	*wasgir*
wasyəɣʷ	*waskar*
wasəɣʷ	*wasga*
was(ə)t(i)	*waskäy*
wasnəɣʷ	*wasgənir*
wasinəɣʷ	*waskänar*
wasənəɣʷ	*wasgäw*

In Kiowa negatives, we find one meaning corresponding to many forms syntagmatically: negation is expressed by a discontinuous marker involving a particle and a suffix. The description of this system is lengthened by the need to specify two forms for one meaning. Many meanings correspond to one form syntagmatically in negatives in Kemant: the meanings of negation, person-number-gender, and aspect are fused in verbal suffixes. A description of this system needs to specify each suffix separately, which increases its length; a small number of generalizations would be enough to describe a more agglutinative system where each meaning is expressed by exactly one form.

Paradigmatic violations of the One-Meaning–One-Form principle can be found in Koyraboro Senni (3) and Karok (4).

(3) Koyraboro Senni (Nilo-Saharan, Songhay) (Heath 1999: 8–9, 57)

a. *n ga koy*
2SG.SUBJ IMPF go
'You are going.'

b. *war si koy*
2PL.SUBJ NEG.IMPF go
'You are not going.'

c. *ay koy*
1SG.SUBJ go
'I went.'

d. *ya na koy*
1SG.SUBJ NEG go
'I didn't go.'

(4) Karok (Karok) (Bright 1957: 67)

a. *kun-iykár-at*
3PL > 3SG-kill-PST
'They killed [him/her].'

b. *pu-ʔiykar-áp-at*
NEG-kill-3PL > 3SG-PST
'They did not kill [him/her].'

Koyraboro Senni has one meaning corresponding to many forms paradigmatically in its system of negation: negation is expressed with different negative constructions using different negative elements with perfective and imperfective aspect. This clearly increases the length of the description as each construction must be separately described. A similar situation is found in Karok, where in addition to the affixes marking negation, different sets of person-number cross-reference markers are used for affirmative

and negative verbs; as a general rule, there are thus two forms for one person-number meaning. However, when second or third person singular subjects act on first person singular objects, the affirmative and negative markers are homophonous, prefix *ná-* in both cases (Bright 1957: 60). This is an example of the fourth possible violation of the principle: many meanings corresponding to one form paradigmatically. This makes the description longer in that many meanings (affirmation vs. negation, second vs. third person acting on first person) have to be specified for one and the same form, including information on which meaning it has in which morphosyntactic and semantic environment.[8]

A few words about the similarity and difference between these two principles and Kusters' and McWhorter's criteria are in order. The principle of Fewer Distinctions resembles McWhorter's Overspecification, but is indifferent as to whether some distinctions go beyond communicative necessity – all that matters is the number of distinctions that a grammar makes within a functional domain. In practice, communicative necessity is very hard to define. How do we decide whether a distinction is or is not necessary for communication? Very few if any grammatical distinctions are necessary in the sense that they would be made in each and every language of the world. McWhorter (cf. above) mentions evidentiality as an example of overspecification; it is true that fewer languages have overt marking of evidentiality than of tense for example, but in the languages that have a highly grammaticalized evidential system, evidential distinctions are important in the network of grammatical meanings that the languages operate with. If needed, meanings expressed by a given rare grammatical category in one language can always be expressed in other languages lexically if not by other grammatical categories. The One-Meaning–One-Form principle corresponds roughly to McWhorter's Structural Elaboration and Irregularity: the more the relationship between meaning and form deviates from the ideal of One-Meaning–One-Form, the more rules are needed in rule-based grammatical models (Structural Elaboration), and irregularities are completely idiosyncratic deviations from One-Meaning–One-Form (Irregularity); the One-Meaning–One-Form principle is more analytic and accounts for a range of phenomena for which McWhorter needs two separate criteria. As to Kusters' principles, my principles of Fewer Distinctions and One-Meaning–One-Form are similar to Economy and Transparency, respectively, but differ from these in their fundamental basis – they are intended in an absolute sense, i.e., as such, independent of (insufficient) psycholinguistic evidence of ease of processing by different types of language users.

The complexity of the system of grammatical meanings that languages distinguish within functional domains is by no means exhaustively accounted for by counting the number of distinctions. The principle of Fewer Distinctions provides a start but many

8. The following clarification may be useful to better understand the many-meanings–one-form situations in terms of description length: I have argued above that *cross-linguistic identification* should be based on meaning/function, not form. This does not mean that linguistic *description* should take this perspective exclusively, only describing how each meaning is formally expressed.

other dimensions need to be taken into account as well. Here I will not enter into a discussion on how such aspects of meaning should be dealt with in terms of complexity, but only mention possible paths to follow. We may for example pay attention to choice-structure relations (Dahl 2004) between meanings distinguished within (or across) functional domains; the use of a given category may determine the availability of other grammatical categories. In Jarawara (Arauan), to take a random example, the choice of past tense enables the speaker to choose between eye-witnessed and non-eye-witnessed evidential categories not available in other tenses (Dixon 2004: 195). Another aspect where the interaction between domains comes in is paradigmatic neutralization of grammatical distinctions in specific environments, e.g., in negation where many languages show loss of grammatical distinctions regularly made in affirmative contexts; in Miestamo (2006b) I have shown how complexity is increased by such paradigmatic restrictions (as well as by other kinds of asymmetries between affirmation and negation, or more generally, between other comparable domains). As to the One-Meaning–One-Form principle, it can cover the whole territory of the complexity of the relationship between meaning and form, but needs to be supplemented with more detailed sub-criteria.

5. Complexity and cross-linguistic rarity

Newmeyer (2007) argues, in a generative perspective, that there is no correlation between grammatical complexity and cross-linguistic rarity. He discusses a number of linguistic phenomena and shows that higher complexity of syntactic derivation (in different versions of generative theory) does not necessarily mean cross-linguistic rarity. Now, in the heart of this issue lies the question what we mean by complexity – how we define it.

When we define complexity in strictly absolute terms, there is no necessary connection between complexity and cross-linguistic rarity. Cost and difficulty (relative complexity), by contrast, are expected to correlate with cross-linguistic frequency: the more costly or difficult a linguistic feature is to process or learn, the less frequent it should be in the world's languages. Cost and difficulty are naturally not the only factors in determining what is common and what is rare.[9] But I would argue, in accordance with Hawkins (2004), that structures that are easy and efficient tend to be preferred in performance and they also find their way more often to grammatical conventions; conversely, difficult and inefficient structures are dispreferred in performance and less

9. The frequency of linguistic features is also affected by accidental statistical properties of the current language population, cf. Maslova (2000). The stability of features is one factor that comes into play here, the frequency of stable features being more likely due to survival from ancestor languages and thus directly linked to the histories of language communities (hegemony relations, survival, death). The frequency of less stable features is more easily affected by functional factors such as cost and difficulty.

often grammaticalized. As we have seen above, (absolute) complexity and cost/difficulty do often go hand in hand, and therefore, in many cases, we could expect increases in absolute complexity to imply cross-linguistic rarity as well. This is however not necessary, and to what extent absolute complexity matches rarity is a separate question.

If we take Newmeyer's approach to complexity as an absolute one, complexity seen as the length of syntactic derivation – length of description – the observation that complexity does not always mean rarity poses no problem. However, to the extent that the syntactic derivations examined by Newmeyer are meant to be psychologically real processing phenomena (cf. the claims of the innateness of UG), such a complexity measure could be seen as an attempt to measure complexity in relative terms, and then the dissociation of complexity and rarity would become problematic given what has been said above about the expected correlation between difficulty and cross-linguistic rarity. My first reaction to this problem is that it casts doubts on the claims of the psychological reality of the theory rather than refuting the proposed correlation between cost/difficulty and cross-linguistic rarity. In a footnote Newmeyer seems to agree with the view advocated here: "[T]ypological generalizations, in particular word order preferences, stem from pressure to reduce parsing complexity" (Newmeyer, 2007, note 9). To some extent at least, the difference thus lies in what is seen as grammar, and what is merely usage in generative theory.

I will now briefly return to the relationship between absolute complexity and cross-linguistic rarity. According to Kusters (this volume), the rule "yes/no questions are related to their affirmative counterpart by a complete reversal of all syllables" would be formally simpler than "a rule that relates these two by assuming complex syntactic structures", but extremely difficult to process, and is therefore not found in any language. Kusters' example may be beside the point in the sense that what his formulation really is, is only a general description of what the rule does, and the rule itself is a very complex recursive mathematical operation. In any case, if one could come up with a large number of genuinely simple rules (or more generally, phenomena with short descriptions) that few or no languages possess, it would be clear that absolute simplicity could not predict much about cross-linguistic frequencies. Be it as it may, if we take an inductive approach and look at what is actually found in the world's languages, and then try to evaluate the absolute complexity of those structures, it is highly likely that there will be some correlation between absolute complexity and cross-linguistic rarity. Perhaps (absolute) simplicity does not always mean ease of processing, but surely (absolute) complexity does in many cases add to processing difficulty.

6. Summary

In this paper I have addressed issues that I consider to be of central importance when language complexity is approached from a cross-linguistic point of view.

In Section 2, I discussed two alternative approaches to complexity – absolute and relative – arguing that for theoretical and methodological reasons, an absolute approach to complexity is to be preferred. In Section 3, I addressed the question of global vs. local complexity concluding that languages can usually only be compared in specific areas of grammar, e.g., in terms of functional domains, and that the concept of global complexity is a problematic one. Section 4 discussed some criteria for studying and comparing languages in terms of complexity, and proposed the principles of Fewer Distinctions and One-Meaning–One-Form as criteria for examining the complexity of functional domains across languages. Finally, Section 5 discussed the relationship between complexity and cross-linguistic rarity from the point of view of the distinctions between (absolute) complexity and cost/difficulty. The focus of the paper has been on theoretical and methodological issues. I have tried to identify some problems encountered in the cross-linguistic study of complexity and clear away some potential misunderstandings by arguing for clear terminological and conceptual divisions. It is my hope that the points I have made here will be of use in future studies of language complexity.

Abbreviations

1	first person
2	second person
3	third person
AFF	affirmative
F	feminine
IMPF	imperfective
M	masculine
NEG	negative/negation
PFV	perfective
PL	plural
PST	past tense
SG	singular
SUBJ	subject

References

Anttila, R. 1972. *An Introduction to Historical and Comparative Linguistics*. New York NY: Macmillan.

Appleyard, D.L. 1975. A descriptive outline of Kemant. *Bulletin of the School of Oriental and African Studies* 38: 316–350.

Bright, W. 1957. *The Karok Language* [University of California Publications in Linguistics 13]. Berkeley CA: University of California Press.

Chomsky, N. 1957. *Syntactic Structures.* The Hague: Mouton.

Croft, W. 2003. *Typology and Universals*. 2nd ed. Cambridge: Cambridge University Press.

Crystal, D. 1997. *The Cambridge Encyclopedia of Language.* 2nd ed. Cambridge: Cambridge University Press.

Dahl, Ö. 2004. *The Growth and Maintenance of Linguistic Complexity* [Studies in Language Companion Series 71]. Amsterdam: John Benjamins.

Dixon, R.M.W. 1997. *The Rise and Fall of Languages.* Cambridge: Cambridge University Press.

Dixon, R.M.W. 2004. *The Jarawara Language of Southern Amazonia*. Oxford: Oxford University Press.

Dryer, M.S. 2005. Genealogical language list. In *The World Atlas of Language Structures,* M. Haspelmath, M.S. Dryer, D. Gil & B. Comrie (eds), 584–644. Oxford: Oxford University Press.

Frajzyngier, Z. with Shay E. 2002. *A Grammar of Hdi* [Mouton Grammar Library 21]. Berlin: Mouton de Gruyter.

Gell-Mann, M. 1994. *The Quark and the Jaguar: Adventures in the Simple and the Complex*. New York NY: W.H. Freeman and Co.

Haspelmath, M. 2006. Against markedness (and what to replace it with). *Journal of Linguistics* 42: 25–70.

Hawkins, J.A. 2004. *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

Heath, J. 1999. *A Grammar of Koyraboro (Koroboro) Senni* [Westafrikanische Studien 19]. Köln: Rüdiger Köppe Verlag.

Hockett, C.F. 1958. *A Course in Modern Linguistics*. New York NY: Macmillan.

Hurd, C. & Hurd, P. 1970. Nasioi verbs. *Oceanic Linguistics* 9: 37–78.

Juola, P. 1998. Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics* 5: 206–213.

Kusters, W. 2003. *Linguistic Complexity, the Influence of Social Change on Verbal Inflection* [Ph.D. Diss., University of Leiden]. Utrecht: LOT.

Li, M. & Vitányi, P. 1997. *An Introduction to Kolmogorov Complexity and its Applications.* 2nd ed. New York NY: Springer.

MacDonald, L. 1990. *A Grammar of Tauya* [Mouton Grammar Library 6]. Berlin: Mouton de Gruyter.

Maslova, E. 2000. A dynamic approach to the verification of distributional universals. *Linguistic Typology* 4: 307–333.

McWhorter, J.H. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology* 5: 125–166.

McWhorter, J.H. 2007. *Language Interrupted: Signs of Non-native Acquisition in Standard Language Grammars*. New York NY: Oxford University Press.

Miestamo, M. 2005. *Standard Negation: The Negation of Declarative Verbal Main Clauses in a Typological Perspective* [Empirical Approaches to Language Typology 31]. Berlin: Mouton de Gruyter.

Miestamo, M. 2006a. On the feasibility of complexity metrics. In *Finest Linguistics. Proceedings of the Annual Finnish and Estonian Conference of Linguistics. Tallinn, May 6–7, 2004* [Publications of the Department of Estonian of Tallinn University 8], K. Kerge and M.-M. Sepper (eds), 11–26. Tallinn: TLÜ.

Miestamo, M. 2006b. On the complexity of standard negation. In *A Man of Measure: Festschrift in Honour of Fred Karlsson on His 60th Birthday* [Special Supplement to SKY Journal of Linguistics 19], M. Suominen, A. Arppe, A. Airola, O. Heinämäki, M. Miestamo, U. Määttä, J. Niemi, K.K. Pitkänen & K. Sinnemäki (eds), 345–356. Turku: The Linguistic Association of Finland.

Newmeyer, F.J. 2007. More complicated and hence, rarer: A look at grammatical complexity and cross-linguistic rarity. In *Phrasal and Clausal Architecture: Syntactic Derivation and Interpretation. In Honor of Joseph E. Emonds* ed. S. Karimi, V. Samiian & W.K. Wilkins (eds), 221–242. Amsterdam: John Benjamins.

Shannon, C.E. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27: 379–423, 623–656. (Reprint with corrections at http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf).

Shosted, R.K. 2006. Correlating complexity: A typological approach. *Linguistic Typology* 10: 1–40.

Stassen, L. 1985. *Comparison and Universal Grammar*. Oxford: Blackwell.

Voort, H. van der. 2004. *A Grammar of Kwaza* [Mouton Grammar Library 29]. Berlin: Mouton de Gruyter.

Watkins, L.J. 1984. *Grammar of Kiowa.* Lincoln NB: University of Nebraska Press.

Complexity trade-offs between the subsystems of language

Gertraud Fenk-Oczlon & August Fenk
University of Klagenfurt

Starting from a view on language as a combinatorial and hierarchically organized system we assumed that a high syllable complexity would favour a high number of syllable types, which in turn would favour a high number of monosyllables. Relevant cross-linguistic correlations based on Menzerath's (1954) data on monosyllables in eight Indo-European languages turned out to be statistically significant. A further attempt was made to conceptualize "semantic complexity" and to relate it to complexity in phonology, word formation, and word order. In English, for instance, the tendency to phonological complexity and monosyllabism is associated with a tendency to homonymy and polysemy, to rigid word order and idiomatic speech. The results are explained by complexity trade-offs between rather than within the subsystems of language.

1. Hierarchy and complexity in the language system

In his famous article on "The architecture of complexity: Hierarchic systems", Herbert A. Simon (1996 [1962]) called the attention of systems theory to hierarchy as a central scheme of organized complex systems:[1]

> Thus my central theme is that complexity frequently takes the form of hierarchy and that hierarchic systems have some common properties independent of their specific content. Hierarchy, I shall argue, is one of the central structural schemes that the architect of complexity uses. (Simon 1996 [1962]: 184)

By a "complex system" he means "one made up of a large number of parts that have many interactions" (pp. 183f). He further states (p. 197) that in

1. We would like to thank the referees and editors for their helpful suggestions.

> a hierarchic system one can distinguish between the interactions *among* subsystems, on the one hand, and interactions *within* subsystems – that is, among the parts of those subsystems – on the other. (Emphasis in the original.)

Later in his article he qualifies the complexity of a structure as critically depending "upon the way in which we describe it" (p. 215).

In Gell-Mann's (1995) conceptualization of *complexity* the description of a structure becomes the crucial point. But he argues that the algorithmic information content (AIC) of such a description is, e.g., because of its context dependency, an inappropriate measure for complexity. "Effective complexity", he says, "refers not to the length of the most concise description of an entity (which is roughly what AIC is), but to the length of a concise description of a set of the entity's regularities." High "effective complexity" in the sense of Gell-Mann amounts to a relatively high number of regularities: it becomes near zero in "something almost entirely random" as well as in "something completely regular, such as a bit string consisting entirely of zeroes." It "can be high only in a region intermediate between total order and complete disorder". In this respect Gell-Mann's conceptualization differs from classical information theory. In information theory as well as in Kolmogorov complexity (cf. Juola 1998), highest complexity is attributed to random order.

A rather recent systems theoretical article by Changizi (2001) studies first of all the development of communication systems. In the abstract of his article he states

> general laws describing how combinatorial systems change as they become more expressive. In particular, . . ., increase in expression complexity (i.e., number of expressions the combinatorial system allows) is achieved, at least in part, by increasing the number of component types. (Changizi 2001: 277)

As an example for the development of "human language over history" he takes English and studies "how the number of word types in a language increases as the number of sentences increases" (p. 281) and concludes that "increasing expressivity in English appears to be achieved exclusively by increasing the number of word types" (p. 283). Notice that in his terminology each "entry in the dictionary is a different word type" (p. 281).

The following criteria of *complexity* may be ordered in the sense of increasing degrees or levels of complexity:

a. The number of components. That there are entities that can form a bigger entity or can be described as components of this bigger entity is the minimum requirement for complexity. But it is a matter of terminology if such a two-step organization – a unit and its elements – already should be considered a minimal hierarchy.
b. The number of components of the components, i.e., the complexity of the components. This criterion refers to a real hierarchy of at least three steps.

c. The number of component types (Changizi 2001). The existence of different types of components makes the entity more complex, irrespective of whether or not it is really hierarchically organized.
d. The number of possible interactions between the components (Simon 1996). The higher a system's complexity with respect to (a), (b), and (c), the higher its complexity with respect to (d).
e. The number of rules determining these interactions, i.e., the number of rules necessary for a concise description of these interactions (Gell-Mann 1995). The higher (a)–(d) are, the higher the possible number of rules determining the interactions within and between the components is.

But if we focus on the system *human language*, we can make out two different types of structuring of this system:

1. Within language, or below the superordinate concept *language*, we may discriminate between different subsystems – or rather levels of description (?) – such as phonology, morphology, syntax, and semantics. This structure is not very distinctive. Semantics, for instance, intrudes on all other subsystems of language.
2. A rather "technical" hierarchy may differentiate, even if ending at the sentence level, between at least five hierarchical steps in the sense of the above criterion b: phonemes, syllables, words, clauses, sentences. The elements of the lowest level are the phonemes, and each unit at a higher level n is, in principle, a complex or a composition of units of the level n – 1 and is in turn an element of units at the level n + 1. But a unit on level n can be identical with a unit on level n – 1, as is the case in monophonemic syllables, monosyllabic words, monoclausal sentences, and, depending on the definition of *clause*, even one-word clauses. Nor should we forget the argument that not all of these divisions are equally clear-cut in any language (c.f. the division of sentences or clauses into words) and at any level: the syllable can – unlike the phoneme (c.f. Ladefoged 2001) – occur as an independent entity and is – unlike the morpheme, for instance – an easily countable component of bigger entities.

Obviously, these two hierarchies (1. and 2.) cannot be fully compatible. But when assigning dimensions of linguistic complexity to different subsystems of language (in Table 1) we encounter some metric parameters belonging to the rather technical hierarchy. The selection of complexity facets in Table 1 is of course "biased" by the findings to be discussed in the chapters below; for instance when taking the number of syllables per word as an indicator of the morphological complexity of these words. Especially in the extremes – monosyllables on the one hand, extremely long words on the other – the n of syllables/word will be an excellent predictor of the words' morphological complexity. From that it follows that languages having rather short words will need more words for encoding a certain semantic unit. These considerations are supported by our cross-linguistic correlation (d) in Section 2.

Table 1. Some dimensions of linguistic complexity related to certain subsystems of language.

Subsystems	Facets of linguistic complexity
phonology	size of phonemic inventory; syllable complexity (= n of phonemes/syllables); n of syllable types
morphology	complexity in word structure (n of morphemes and n of syllables per word); n of morphological cases, gender distinctions etc.; opaqueness of morphological forms
syntax	rigid (?) word order; hypotactic constructions
semantics	n of meanings per expression (homonymy, polysemy)

2. Cross-linguistic correlations

2.1 Previous results

A previous study by the authors (Fenk and Fenk-Oczlon 1993) revealed a set of significant and mutually dependent cross-linguistic correlations between the four variables number of phonemes per syllable (syllable complexity), number of syllables per word, number of syllables per clause, and number of words per clause:[2]

a. The more syllables per word, the fewer phonemes per syllable.
b. The fewer phonemes per syllable, the more syllables per clause.
c. The more syllables per clause, the more syllables per word.
d. The more syllables per word, the fewer words per clause.

Correlation (a) is a cross-linguistic version of a law originally found by Menzerath (1954: 100) in German: "The relative number of sounds decreases with an increasing number of syllables [per word]" (our translation). Additional calculations admitting higher order (quadratic, cubic, logarithmic) functions resulted – for obvious reasons – in higher determination coefficients (Fenk & Fenk-Oczlon 1993, e.g., 18f) than the

2. The main database for this statistical reanalysis originates from a quasi-experimental study by Fenk-Oczlon (1983): native speakers of 27 typologically different languages were instructed to translate a set of 22 German "mono-clausal" sentences into their mother tongue and to determine the n of syllables of each of the sentences produced. The written translations (in facsimile on the pages 104–182 of this study) allowed one to enumerate the n of words per sentence. The number of phonemes was determined with the help of the native speakers and grammars of the respective languages. The experiments were continued with 29 languages in our (1993)-study and 34 languages in the (1999)-study. Due to research grants (Fulbright and University of Klagenfurt) the size of the sample will arrive at 65 in the near future.

linear correlations. The whole set of correlations (a)–(d) was confirmed in a later study (Fenk-Oczlon & Fenk 1999) with an extended sample of 34 languages, 18 Indo-European and 16 non-Indo-European.

In the case of Menzerath's law on the single-language level, the best fit (e.g., a determination coefficient of .995 for German) could be achieved when using the model of exponential decay in order to describe the syllable complexity (n of phonemes per syllable) as a function of the number of syllables per word (Fenk, Fenk-Oczlon & Fenk 2006). That this function of an exponential decay can be shown (same article, p. 332) as a special case of Altmann's (1980) mathematical generalization of Menzerath's original law is not that surprising. Altmann's law, often referred to as the "Menzerath-Altmann law" or even, as already in Altmann (1980), as "Menzerath's law", is so general that Meyer (2007) could specify a mere stochastic mechanism generating such relations "in ensembles of hierarchically structured entities of whatever kind."

Summarizing the results of our previous work we may say two things. Firstly, as a plausible consequence of the negative correlations of the syllable complexity (A) with both the number of syllables per word (B) and the number of syllables per sentence (C), we expected and indeed found a positive correlation between B and C (Fenk & Fenk-Oczlon 1993: 18).[3] This is one of several arguments for the mutual dependency of the correlations and for a systemic view of language variation. Our cross-linguistic correlations can best be understood as balancing effects between the subsystems of language. Such balancing effects seem to provide a cross-linguistically rather constant size (duration) of clauses and of mono-clausal sentences.

Secondly, syllable complexity seems to play a key role within this system (Table 2). It interacts, first of all, with other metric properties. (In each column of Table 2 the first four rows including the headings "paraphrase" the correlations (a)–(d) mentioned above.) Syllable complexity is also significantly associated with word order (Fenk-Oczlon & Fenk 1999: 163f) and almost significantly with the number of cases, which is in turn significantly associated with adposition order (Fenk-Oczlon & Fenk 2005: 81). These two studies offer some further statistical as well as theoretical arguments concerning the last three rows in Table 2. And syllable complexity also interacts, more or less directly, with semantics, as we will argue in Section 3.

---

3. In the case of two given correlations r_{xy} and r_{xz} with the same sign (a positive sign in both correlations or a negative sign in both correlations) a third "correlation" between Y and Z, i.e., any coefficient r_{yz} different from zero, will rather show a positive sign. In cases of different signs in the two given correlations we rather have to expect a negative sign in the third correlation. The higher the correlations r_{xy} and r_{xz}, and the higher therefore the determination coefficients r^2 and the variance explained by them, the higher the plausibility of the assumption of a correlation r_{yz}. For a more detailed discussion of this sort of statistical reasoning see Fenk-Oczlon & Fenk (2005) and Fenk & Fenk-Oczlon (2006).

Table 2. Associations between syllable complexity and some other metric and non-metric properties.

High syllable complexity	Low syllable complexity
low n of syllables per word	high n of syllables per word
low n of syllables per clause	high n of syllables per clause
high n of words per clause	low n of words per clause
VO order	OV order
low n of morphological cases	high n of morphological cases
prepositions	postpositions
cumulative case exponents	separatist case exponents
stress timed	syllable timed
fusional or isolating morphology	agglutinative morphology

2.2 New assumptions

A high syllable complexity in a certain language requires a rather large phonemic inventory. (A high syllable complexity can only be achieved by large initial and final consonant clusters. In languages showing comparable degrees of freedom in the combinatorial possibilities of consonants, those having a larger inventory of consonants will incline to bigger consonant clusters.) A relatively high syllable complexity is in turn a precondition for a high variability of syllable complexity and therefore also for a high number of syllable types; a high number of syllable types will not be possible in a language with a maximum of, let us say, two phonemes per syllable. And a high number of syllable types is a precondition for a large inventory of monosyllabic words. Facing such implications of combinatorial possibilities we have to be aware of two methodologically relevant aspects.

On the one hand it is plausible to assume that a certain system realizes a good deal of its combinatorial possibilities. On the other hand, the concrete system will hardly ever realize the total of cogitable possibilities. In the system "language", as we know, the phonotactic possibilities realized fall below the number of combinatorial possibilities. Thus, one cannot mathematically determine a language's mean number of phonemes per syllable from a given size of its phonemic inventory or its number of syllable types from a given mean or maximum syllable complexity or its number of monosyllables from a given number of syllable types. Instead, we always depend on empirical investigation if we want to determine the extent to which combinatorial possibilities are realized.

Thus our chain of preconditions and requirements (first paragraph in this section) should be reformulated in the sense of a chain of statistically testable tendencies. The data regarding syllable complexity (X), the number of syllable types (Y) and the number of monosyllabic words (Z) could be put together from Menzerath's (1954) descriptions of eight different languages or could be calculated from these data. This sample of only eight Indo-European languages of Europe of course cannot claim to be

representative for the total "population" of natural languages. And it should be noted here that the values regarding (X) and (Y) are the values *realized in the monosyllabic words* of the respective languages. But if it is true that especially monosyllables "use" a considerable range of the spectrum of syllable types, or – in rather analytic/fusional languages – almost this whole spectrum, then these values are acceptable at least as good indicators for a language's syllable complexity and its total number of syllable types.[4] Thus, the following hypotheses can be evaluated:

- Hypothesis I: the higher a language's number of syllable types, the higher its number of monosyllabic words.
- Hypothesis II: the higher a language's syllable complexity, the higher its number of syllable types.
- Hypothesis III: the higher a language's syllable complexity, the higher its number of monosyllabic words.

In a second step we tried to find data generated by one author and one method regarding the size of the phonemic inventory (W) in the given sample of eight languages. Determining the number of phonemes is generally problematic (Bett 1999), and one of the most problematic areas is the analysis of phonetic diphthongs (Maddieson 1984: 161). Spanish, for example, seems to have no diphthongs in Ladefoged (2001) but has five diphthongs according to Campbell (1991). Campbell offers exact numbers regarding consonants and monophthongs in our eight languages but remains incomplete as to the number of diphthongs. Thus, in testing the first chain in our above arguments (first sentence in this section) we decided to take the sum of consonants and monophthongs as a value indicating the size of the phonemic inventory:

- Hypothesis IV: the bigger a language's phonemic inventory, the higher its syllable complexity.

2.3 Evaluation and results

Our hypotheses do not imply any specific assumption regarding the shape of the respective regressional functions, so that the standard correlation (Pearson's product-moment correlation) and the corresponding test of significance was the first choice. If this application of the linear model reveals significant coefficients, this means a confirmation of the hypotheses in a very rigid examination, because further appropriate

4. In languages showing a high proportion of monosyllables we encounter, apart from monosyllabic function words, also a high number of monosyllabic content words. Especially languages with such a manifest tendency to monosyllabism show a tendency to isolating techniques (not necessarily the other way round), high syllable complexity and high variability of syllable types.

curve-fitting procedures will necessarily result in higher, never in lower determination coefficients. In the following we will compare the results of the application of the linear model with the results of the data-driven search for better fitting models. From those "simple" (c.f. Mulaik 2001) models tested the model of exponential growth achieved the best fit in all of the following evaluations (see Figures 1–3).

Hypothesis I predicts a cross-linguistic correlation between the number of syllable types and the number of monosyllables. Or, to put it the other way round: a large inventory of monosyllabic words will go hand in hand with a large inventory of different syllable types: V; CV, VC; CCV, CVC, VCC; . . . CCVCCCC; . . . CCCCVCCCC. This hypothesis is in line with the argument that, in the system "language", complexity trade-offs will happen between rather than within its subsystems. A high number of syllable types reflects a high (variation of) phonological complexity while a high number of monosyllables reflects a low complexity in word formation – a low complexity in terms of n of syllables and, indirectly, in terms of morphemes as well.

Table 3. The frequency of monosyllables consisting of different numbers (1, 2, . . . 8) of phonemes (Data assembled from Menzerath 1954).

	n of phonemes per monosyllable	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	English	14	326	2316	2830	1199	161	8	2
2	German	9	114	645	962	444	69	2	
3	Romanian	8	81	480	474	135	19	1	
4	Croatian	6	42	353	273	42	1		
5	Catalan	11	94	285	265	25			
6	Portuguese	9	84	177	51	4			
7	Spanish	10	59	115	70	9			
8	Italian	4	50	32	7				

Table 4. Data collected or calculated from Menzerath's descriptions of eight different languages.

		X		**Y**	**Z**
		n of phonemes per monosyllable		**n of syllable types realized in monosyllables**	**n of monosyllables**
		X_{mean}	X_{max}		
1	English	3.787	8	43	6856
2	German	3.861	7	35	2245
3	Romanian	3.591	7	16	1198
4	Croation	3.427	6	12	717
5	Catalan	3.293	5	11	680
6	Portuguese	2.868	5	9	325
7	Spanish	3.034	5	17	263
8	Italian	2.452	4	8	93

Table 3 assembles relevant data from Menzerath's typological descriptions of eight different languages (Menzerath 1954: 112–121) regarding the frequency of monosyllables of different complexity. From the rows in Table 3 one can calculate the mean n of phonemes per monosyllable in any single language. Taking Italian as an example: 1 × 4 + 2 × 50 + 3 × 32 + 4 × 7 = 228 divided by the cumulative frequencies (4 + 50 + 32 + 7 =) 93 is 2.452. Table 4 compares this mean syllable complexity (X_{mean}) with other data by Menzerath. It allows a direct test of our Hypotheses I–III in the sense of cross-linguistic correlations. The first result was a clear confirmation of Hypothesis I: *the more syllable types (in monosyllables), the more monosyllables,* r_{yz} = + .895 ($p < .01$).[5] This means a determination coefficient of .801. The determination coefficient achieved with an exponential growth model (Figure 1) is .978. Figure 1 also highlights the essential contribution of the two Germanic languages (English and German, cf. Table 4) to that regression.

Hypotheses II and III both concern the parameter of syllable complexity. The examination of Hypothesis II again revealed significant correlations:

– Hypothesis II: the higher the mean number of phonemes per syllable, the more syllable types, r_{xy} = + .76 ($p < .05$)

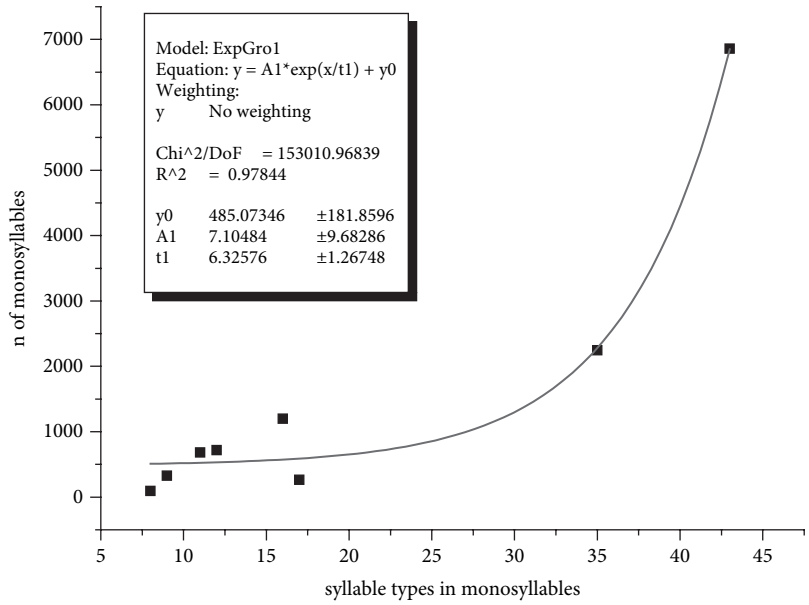


Figure 1. The regression regarding Hypothesis I with the exponential growth model.

5. This significant result could already be presented at the 2005 Helsinki Symposium on Approaches to Complexity in Language.

– Hypothesis II': the higher the maximum number of phonemes per syllable, the higher the number of syllable types, r_{xy} = +. 835 (p < .01).

This means a determination coefficient of .578 or .697, respectively. The determination coefficients achieved with a model of exponential growth (Figure 2) were .820 or .801. This means that in the seemingly most appropriate regressional model the mean syllable complexity was a better predictor of the number of syllable types than the maximum syllable complexity.

Because of different phonotactical possibilities in the respective languages, such correlations cannot be predicted or explained by mere combinatorial effects. In the case of a trivial relation between the mean number of phonemes per monosyllable (X_{mean} in Table 4) and the number of syllable types (Y) a regression r_{xy} should be "perfect"; German for instance should have, as compared with English, either a lower value in X_{mean} or a higher value in Y. The very same can be said if we take, instead of the *mean* number of phonemes, the *maximum* number (X_{max}) of phonemes: in the case of a trivial relation it would not be possible to find three Romance languages with an equal maximum of five phonemes per syllable showing different numbers of syllable types ranging from nine in Portuguese to 17 in Spanish, or to find two languages (Romanian, German) with an equal maximum of seven phonemes per syllable but 16 vs. 35 syllable types (X_{max} in Table 4 and right panel in Figure 2).

The inferential step from syllable complexity (X) to the number of monosyllabic words (Z) is not trivial either: of those three languages with an equal value for maximum syllable complexity (X_{max} in Table 4), Spanish has indeed by far the highest number of syllable types (17) but by far the lowest number of monosyllables: 263 as compared with 680 in Catalan. Notably, Catalan also shows the highest syllable complexity of those three languages. The relevant Hypothesis III is not only the remaining link in our chain of arguments. It was, moreover, inspired by our significant correlation (a) in Section 2.1:

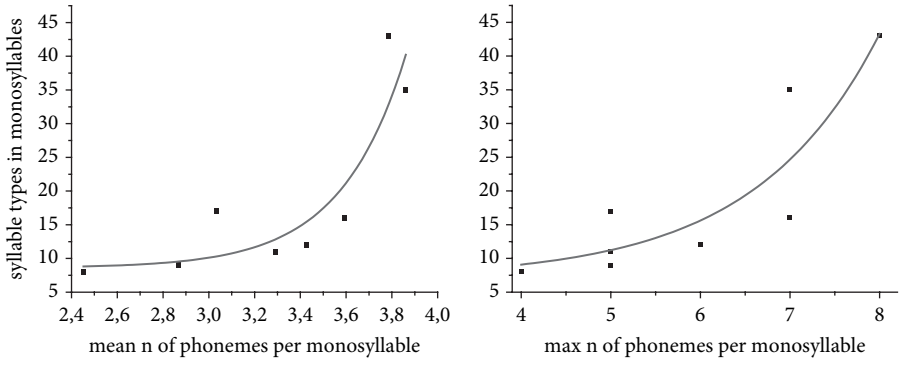


Figure 2. Regressions regarding Hypothesis II. The independent variables are mean syllable complexity (left panel) and maximum syllable complexity (right panel).

"The higher the number of phonemes per syllable, the lower the number of syllables per word". The absolutely lowest number of syllables is of course realized in the monosyllabic word so that there is only a small step to predicting an association between high syllable complexity and a strong tendency to produce monosyllabic words.

The result of the statistical examination was an almost significant correlation when the n of monosyllables was correlated with the mean syllable complexity and a significant correlation when correlated with the maximum syllable complexity:

- Hypothesis III: the higher the mean number of phonemes per syllable, the more monosyllabic words, $r_{xz} = + .64$ ($p < .1$)
- Hypothesis III': the higher the maximum number of phonemes per syllable, the higher the number of monosyllables, $r_{xz} = + .807$ ($p < .05$).

This means a determination coefficient of .410 or .651 respectively. The determination coefficients achieved with a model of exponential growth (Figure 3) were .536 or .980.

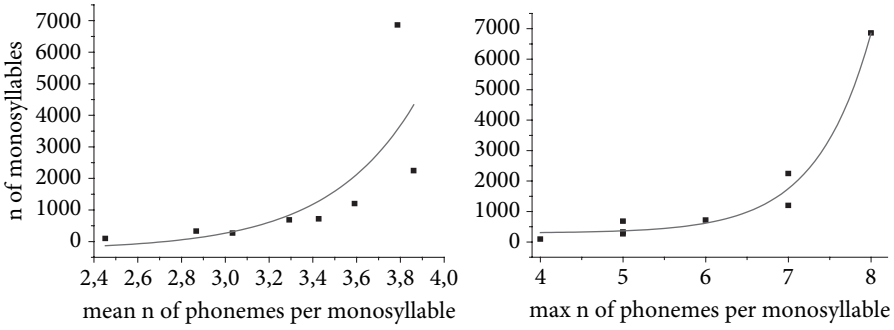


Figure 3. The regressions regarding Hypothesis III, with Xmean (left panel) and Xmax (right panel) as the independent variables.

In the evaluation of the remaining Hypothesis IV we counted the numbers of phonemes, excluding diphthongs, in our eight languages as reported by Campbell (1991): English 35, German 40, Romanian 28, Croatian 34, Catalan 30, Portuguese 31, Spanish 25, and Italian 29. The correlation of these numbers with mean syllable complexity (X_{mean} in Table 4) was not too far from being significant and showed, as expected, a positive sign: *the bigger the phonemic inventory, the more phonemes per syllable,* $r_{wx} = + .622$ ($p < .1$). This means a determination coefficient of .387. The determination coefficient achieved with the exponential growth model was .411. When the syllable complexity figured as the independent variable, this determination coefficient was .650. This might be seen as a further indication (cf. Fenk-Oczlon & Fenk 2005, Fenk & Fenk-Oczlon 2006) for the central role of syllable complexity in language variation.

The linear correlations of the segmental inventory (W) with the other variables in Table 4 (X_{max}, Y, and Z) showed positive signs as well: + .573, + .635 *($p < .1$)*, and + .508.

2.4 Discussion

What we could show significantly in a small sample of eight Indo-European (one Slavic, two Germanic, five Romance) languages is a new set of mutually dependent, cross-linguistic correlations between the number of monosyllables, the number of syllable types (in monosyllables), and syllable complexity (in monosyllables). All the correlations of the phonemic inventory size with other variables showed the expected signs, and two of them were almost significant.

We are optimistic that this result concerning mutually dependent parameters will hold for the monosyllabic system in general. But we cannot predict if our generalizations transcending the monosyllabic system will hold when, in days to come, reliable and comparable characteristic values regarding the syllable complexity and the number of syllable types are, disconnected from the monosyllabic system, available for a more widespread sample of languages.

In the case of a negative correlation between e.g., syllable complexity and number of syllables per clause one may easily identify a balancing mechanism. But how can we state such a balancing effect in view of our three positive correlations? Our answer to this question is that both a high number of phonemes per syllable and a high number of syllable types mean or reflect high phonological complexity, while the tendency to produce many monosyllables reflects low complexity in word structure.

2.4.1 *A short excursion into diachrony*

If we consider the repertoire of monosyllables as the system in question and the syllable types as its component types and take a look at the diachronic changes of English from this perspective, we find developments that conform to both Changizi's claims (see Section 1) and our model of complexity trade-offs: a comparison of the Beowulf Prologue in Old English (OE) with its translation into Modern English (ME) shows a remarkable increase of monosyllables from 105 in OE to 312 in ME and a concomitant increase of the mean syllable complexity from 2.63 phonemes in OE to 2.88 in ME.

We could not determine the exact number of syllable types in OE, but as we saw in Section 2.3, high syllable complexity favours a high number of syllable types (our significant correlations II and II'). Other indications for a higher number of syllable types in ME than in OE are: the loss of final segments as in *guest* versus *gas-tir*, *horn* versus *hor-na* (examples from Lehmann 1978) resulted in an increase of final consonant clusters and therefore also of syllable types. And while the initial consonant clusters seem to be similar in OE and ME, Modern English shows a higher number of complex final clusters such as *strands* /strændz/ CCCVCCC, *glimpsed* /glimpst/ CCVCCCC.

2.4.2 *Where the trade-offs happen and why they do not indicate an equal overall complexity of languages*

All of our previous and present results indicate that in the system "language" one meets complexity trade-offs between rather than within the subsystems, and that within a subsystem one may even observe a diachronic increase in many parameters of complexity.

This corresponds to the results reported by Maddieson (1984) concerning phonology. He showed convincingly that languages with a large consonant inventory also tend to have a large vowel inventory (p. 17). A lower number of manner contrasts was not found to be compensated by a higher number of place contrasts of stops and fricatives (p. 18), and the often stated assumption that a small phonemic inventory is compensated by more complex suprasegmentals (i.e., tone, stress) could not be confirmed either. Maddieson's analysis of 56 languages showed on the contrary that languages with simpler segmental inventories tend to have less elaborated suprasegmental properties (p. 21). He also mentions a positive though rather weak and insignificant correlation between segmental inventory size and syllable inventory size. In a recent study (Maddieson 2006: 110f) with an extended sample of languages he found a significant difference: the most consonants occurred in the inventory of languages with complex syllable structures, the least in those with simple syllable structures. Our positive correlation between phonemic inventory and syllable complexity is, though not significant, in line with these findings by Maddieson.

Menzerath's (1954) law points to complexity trade-offs on the intra-language level, and our results (Fenk & Fenk-Oczlon 1993 and present study) point to such trade-offs in cross-linguistic comparison. Such cross-linguistic trade-offs or balancing effects gave rise to the attractive idea of something like an equal overall complexity in all our natural languages. We agree with the arguments by e.g., Miestamo (this volume) that there is no possibility to verify such a hypothetical equality and would like to stress the fact that this idea of equality is in no way supported by those correlations pointing to balancing effects. Let us illustrate this with an example with only two parameters:

A low-budget University department regularly records the number of printouts and copies per individual member. (Some are proud of their "productivity", others of their "economy".) A "cross-subject" negative correlation between the number of printouts and copies indicates a balancing effect: the more copies, the fewer printouts, and vice versa. But this correlation does not at all mean that all the members of the department achieve the same sum of copies plus printouts. It is fully compatible with some members producing both far more printouts and far more copies than others. We may conclude: not even very clear cross-linguistic balancing effects can be interpreted in the sense of an equal overall complexity in the respective languages. And in language we have not only two and very clearly defined parameters, but different subsystems whose complexity is measured by different, more or less well defined parameters.

3. Conceptualizations of "semantic complexity" and a look at pidgin languages

Pidgin and creole languages are often supposed to be the world's simplest languages (e.g., McWhorter 2001): they show a small lexicon and a low complexity in phonology, morphology, and syntax. We will argue that this is to some degree compensated for by

semantic richness of the expressions, i.e., by what we call high "semantic complexity". According to this conceptualization, a large proportion of expressions encoding a large repertoire of different meanings, i.e., a tendency to homonymy and polysemy, may be regarded as indicating high "semantic complexity".

But we have to admit that such a concept of *semantic complexity* goes beyond the scope of a rather technical concept of hierarchical steps of complexity. The different meanings that can be assigned to a certain verbal expression can hardly be viewed as the parts of this expression, nor can the expression be viewed, strictly speaking, as a complex of its possible meanings.[6] And relevant expressions need not in any case be viewed as "poly-functional", i.e., as having "inherently" many meanings that are triggered in certain contexts: Gil (this volume) observed a high availability of apparently associational interpretations of expressions in isolating languages with basic SVO word order. He views the prominence of such an "associational semantics" as a characteristic of simplicity rather than complexity. But the scope of perspectives on semantic complexity seems to be broader (Raukko 2006).

What are the alternatives if one denies "semantic complexity" as operationalized above? Maybe it is generally misleading to talk about *meaning(s)* as something that can be ascribed to an isolated expression instead of something that comes about by the context and context-dependent associations. The last two of the following conceptualizations try to cope with this problem:

1. If one takes a "word" as a unit of a particular sound pattern plus a particular meaning, then the number of words has to be identified – at least in face of the unrelated meanings in homonyms – with the number of different meanings instead of the number of different sound patterns. This would mean a relatively large and highly variable adaptive lexicon in pidgin languages instead of a restricted lexicon.
2. We relate "semantic complexity" to Simon's (1996) complexity criterion "number of interactions between components". Taking the word or the lexical morpheme as the component, it may be argued that the "different meanings" of this component come about by those linguistic contexts admitting the occurrence of this certain component.[7] These linguistic contexts represent the possible interactions with other components.

---

6. In this respect we fully agree with Meyer (in press). He questions (in footnote 8) several applications (e.g., in Cramer 2005) of Menzerath's law: "Can the different meanings of a polysemous lexeme really be treated as 'constituents' of the lexeme?"

7. This conceptualization is inspired by ideas of Wittgenstein, the others are more or less in line with constructivistic or otherwise mentalistic conceptualizations. A separate article would be necessary to investigate such attributions.

3. We say that "semantic complexity" is not an inherent property of the external symbol system "language" but has to be allocated "in the heads" of its users.[8] The efficient use of a pidgin language demands, even more than a standardized and highly "overlearned" language, high context sensitivity, awareness of the situational context as well as intuitive and fast associative checks and decisions. Using a pidgin language is more like sailing than driving a heavy motorboat.

And while it may be possible to count or to estimate the number of meanings per expression in highly standardized languages with their large volume lexica and canonical assignments of literal and figural meanings to a certain expression, this will be almost impossible in languages that incessantly produce new meanings and more or less metaphorical applications of their verbal expressions.

Pidgin languages show a tendency to reduce the phonological complexity of expressions from both the substratum language and the superstratum language, a tendency that often results in homophony. Todd and Mühlhäusler (1978: 11) report of an example: in some idiolects of Cameroon Pidgin the word *hat* has acquired four different meanings. This homophone originates from a simplification of English phonemes and syllable structures. It is the result from a sound merger of the English words *heart*, *hot*, *hurt*, and *hat*.

Pidgin languages have a rather small vocabulary from the very beginning, and amalgamations of the sort illustrated above are likely to keep the small lexicon small – at least at the sound level (see alternative 1 above)! The single entries into a small lexicon generally incline to both a frequent use and polysemy. (The association between frequency and polysemy is a well known phenomenon since Zipf 1949). This inclination to homonymy and polysemy and to an accumulation of more or less "figural meanings" goes hand in hand with or is favoured by the creation of non-conventionalized ad-hoc metaphors and by the adoption of idioms from both the superstratum and the substratum language.

Polysemy and especially homonymy, because of the semantically more distant values associated with one word (Raukko 2006: 358), may be regarded as contributing to "semantic complexity": in order to be effective, the language counterbalances its simplicity in the lexicon with complexity in semantics, i.e., in a creative and flexible accumulation of context-specific meanings. Viewing our alternative possible operationalizations one might allocate complexity not in the external symbol system but "in the heads" of the users and in their "mental navigation" between the cultural backgrounds of superstratum and substratum language. All this means higher cognitive costs, in particular if one assumes that homonyms and polysemous words have to be stored and memorized together with different possible contexts. And it would boil down

8. According to Evans (2005: 34) "semantic structure derives from and mirrors conceptual structure . . . Hence linguistic polysemy reflects complexity at the level of mental representation".

to compensatory effects not between or within subsystems of language but, instead, between the external language system and internal, i.e., mental representations.

Pidgin languages are languages in statu nascendi and do not show the "overall complexity" of more developed languages. Nevertheless, we can state at least two balancing effects within pidgin languages. Firstly, the relatively small lexicon and the low complexity in phonology go hand in hand with a high "semantic complexity", i.e., with a tendency to homonymy, polysemy, and non-conventionalized metaphors. Semantic complexity in this sense will grow tremendously in more or less conventionalized expressions of pidgin languages reflecting extremely different cultural backgrounds. The following examples (1) to (3) by Todd and Mühlhäusler (1978) may illustrate this:

Cameroon Pidgin (p. 11):

(1) *Wash bele*
wash belly
'the last child'

Tok Pisin (pp. 24f):

(2) *Han bilongan I nogut*
Hand POSS[9] is not.good
'she is menstruating'

(3) *karim lek*
to.carry legs
'a form of courtship in the new Guinea highlands'

Secondly, the relatively small lexicon and the low complexity in phonology go hand in hand with a higher complexity of words in terms of n of syllables. As compared with English, which is the superstratum of many pidgin languages, those pidgin languages exhibit a higher proportion of bisyllabic words (Hall 1966; Heine 1973). This tendency to bisyllabic words means, unlike an increase in "semantic complexity", a higher complexity in a rather technical sense of the term. As syllable structures are simplified, the number of syllables per word increases. (See correlation (a) between n of phonemes per syllable and n of syllables per word in Section 2.)

4. Low complexity in word structure – high semantic complexity and idiomatic speech?

How can semantic complexity, as conceptualized above, be localized within the language system? We think that low morphological complexity especially in word

9. The abbreviation POSS means possessive,

structure favours a higher semantic complexity – e.g., a tendency to homonymous and polysemous expressions encoding a higher number of senses – which in turn requires and favours a rather formulaic speech, i.e., stable fragments of speech that allow a quick identification of the context-relevant meaning. Therefore the context should be stored and memorized together with the homonymous and polysemous words. A high proportion of idioms and a tendency to formulaic speech increase the cognitive costs in the acquisition of the respective language. As to these new assumptions we will refer to statistical information supporting our arguments but cannot offer cross-linguistic *correlations* in the sense of inferential statistics. Instead we will substantiate these assumptions by taking English as an example and by a contrastive comparison between English and Russian.

4.1 English as an example

English has a high number of monosyllabic words, roughly 8000 according to Jespersen (1933). A high number of very short, i.e., monosyllabic, words is associated with a high number of syllable types (our correlation I) and with a large phonemic inventory. All these are indications pointing to a high phonological complexity in English. Monosyllabic words, on the other hand, are not suitable for coding many morphological categories. This amounts to a low morphological complexity.

Languages with a high number of monosyllabic words tend, moreover, to have a higher number of homonyms. According to Jespersen (1933) there are about four times more monosyllabic than polysyllabic homonyms: "The shorter the word, the more likely is it to find another word of accidentally the same sound". Homonymy affects of course also "parts of speech" and most of the "grammatical homophones" such as *love* (verb and noun) or *round* (noun, adjective, adverb, preposition, and verb) are again monosyllables. Because of the well known association between frequency and polysemy on the one hand and frequency and shortness on the other, polysemy should also be a frequent phenomenon in monosyllabic words. Both homonymy and polysemy may be viewed, as already mentioned, as dimensions of semantic complexity.

4.1.1 *English phrasal verbs – monosyllabic and idiomatic!*

Phrasal verbs may be considered as a special case of idioms. They consist of a verb in connection with an adverb and/or preposition such as *get by, get along with, get in, get over, act on, act up.* Phrasal verbs are idiomatic, because their meaning cannot be derived from the meaning of each word separately. The verb as well as the adverb or preposition forming the phrasal verb are often polysemous. But in combination they are quite unambiguous, despite the fact that the phrasal verb may again have more than one idiomatic meaning, as e.g., in *go for* and *set off.*

In a short analysis of a collection of 1406 English phrasal verbs we found that 1367 or 97% of the verbs that were part of the phrasal verb construction were monosyllabic

(39 phrasal verbs included a bisyllabic verb and only one was found with a trisyllabic verb.).[10]

Phrasal verbs have to be memorized holistically. One has to know rules concerning their separability, e.g., *add up* (separable) versus *get around* (inseparable). In case of separability one has to know in addition that a pronominal direct object must always be put between verb and preposition.

4.2 A short comparison between English and Russian

The tendency to short words, to polysemy and homonymy and to idiomatic speech becomes a distinct characteristic of English especially in comparison with Russian. Relevant data (see Table 5) counted and reported by Polikarpov (1997) perfectly match with our model.

Table 5. Comparing Russian and English (data by Polikarpov 1997).

Word length	Russian words are on average 1.4 times longer than English words
Polysemy	English words have on average 2.7 meanings, Russian words only 1.7
Homonyms	in English at least 2000, in Russian about 500
Idioms	in English roughly 30,000, in Russian about 10,000

Do the tendencies to idiomatic speech and to rigid word order in general, mean a higher or a lower complexity? This is the central question to be discussed in the following section.

5. Final discussion

The tendency to idiomatic speech marked the endpoint of several of our chains of arguments so far. Taking English as an example one finds: high phonological complexity – low morphological complexity – high semantic complexity – rigid word order and idiomatic speech. What we have avoided so far, because of conflicting positions between the first and second author, is an answer to the question whether rigid word order and idiomatic speech indicate high or low complexity. Let us start with the first author's position that is illustrated in Figure 4:

High phonological complexity, e.g., a large number of syllable types, is associated with a tendency to monosyllabism. Monosyllabic words are not suitable for coding many grammatical morphemes (i.e., isolating morphology); this means low morphological complexity and a low number of morphological rules. Furthermore, monosyllabism is strongly associated with lexical and part-of-speech ambiguity. This means high semantic

10. The data was obtained from http://usingenglish.com, 22.02.2006.

complexity. To keep the language system efficient, resolving grammatical ambiguity requires or favours rigid word order. Lexical ambiguity requires collocations for resolving homonymy and polysemy, etc. All this results in a higher word order complexity: more word order rules, more lexical collocation rules, formulaic speech, and idioms such as e.g., phrasal verbs.

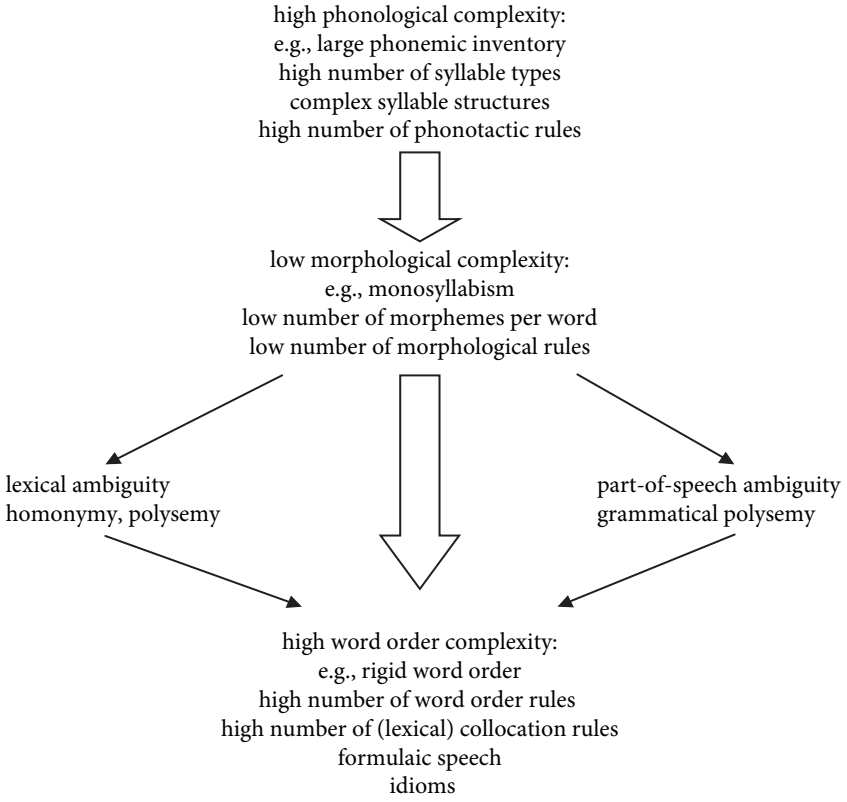


Figure 4. Complexity trade-offs between the subsystems phonology, morphology, semantics, and syntax.

Information theory, however, offers a somewhat different approach. This approach is, in the opinion of the second author, fully compatible with the descriptions of balancing effects in the earlier sections of this paper. All the following properties contribute to "rigid word order" in a more general sense: formulaic speech, including idiomatic speech and phrases, as well as "rigid word order" in a more specific sense, i.e., in the sense of those rules we know e.g., from English: the verb always comes after the subject; in questions one has to put the auxiliary before the subject, etc.

Most linguists would attribute high complexity to rigid word order, at least to rigid word order in the more specific sense of the term. But according to the "logic" of the

balancing effects hypothesized above it should be associated with low complexity: the meaning of the individual words that constitute an idiom differs from the dictionary definitions of these words and depends on the meaning of the whole of which it is a part. In other words: this whole is a "prefabricated" (Wray and Perkins 2000: 1) series of individual words and the contribution of the individual word is specified by this series. Given a sufficient familiarity of the idiom, no other meanings will be associated with the individual words: the prefabricated linguistic context effectively "selects" or specifies the meaning within the phrase. Thus we may say that the high redundancy – or low complexity? – of highly overlearned idioms is a very selective filter that allows high complexity in word semantics.

This view is actually consistent with information theory. Rigid word order boils down to high redundancy, high predictability, and low informational content. Given a fragment of a rather redundant series it is relatively easy to anticipate or to reconstruct the whole. The number of errors that are made by native speakers in the guessing game technique (Shannon 1951) will be rather low in redundant series. This technique can only measure, though this was not intended by Shannon, the subjective information, i.e., the information a text contains for the specific guessing subjects. A text that is highly redundant (for a highly competent speaker) is without any doubt simple (for the highly competent speaker), and not at all complex.

A low level of complexity of the sequence of elements and "supersigns" may come about by a very low number of rules. Very few and very simple rules (e.g., exclusively CV-syllables, bisyllabic words and VO order!) may lead to an extremely high redundancy of the string. In such a language we will observe low complexity in both the string and the production rules (the "source"). But a text of a comparably high redundancy and low complexity may also be produced by a huge number of less restrictive rules. In such a language the text remains simple and redundant for the highly competent speaker, but the "production system" may be regarded as highly complex according to Gell-Mann's criterion "number of rules". Complexity in this sense will show in the process of language acquisition, especially in Second Language Acquisition (SLA), and in the attempts of linguists to extract the phonological, morphological and syntactical rules of the respective language. A third virtual language may also have a huge number of rules, being highly complex in this sense, but these rules are in part less rigorous, more concerned with aesthetic and pragmatic principles that allow considerable freedom in sentence construction, while other rules of this language may be rigorous but only of "local" effectiveness. (Higher freedom in word order does not mean randomness and does not necessarily mean a lower number of rules determining word order. This virtual language requires, on the contrary, many additional "stylistic" rules that decide in cases of conflicts.) Such a language is complex in both the production rules and the "surface".

From this point of view, the trade-off is between "semantic complexity" on the one hand and "word order complexity" on the surface on the other, independent of the number of rules contributing to word order complexity: high complexity in word semantics requires or favours low complexity (high rigidity, high redundancy) in word order and

vice versa. This would suggest the following modification of the succession illustrated in Figure 4: high phonological complexity – low morphological complexity – high semantic complexity - *low* complexity (high rigidity) in word order.

6. Concluding remarks

Functional explanations of all the complexity trade-offs reported or hypothesized above refer to economy principles in language use. The essence of such explanations (Fenk-Oczlon & Fenk 1999, 2002) is a tendency of natural languages to keep the size of clauses and the information flow within these clauses rather constant. This tendency forces complexity trade-offs between those units (syllables, monosyllables and polysyllabic words, clauses and "mono-clausal" sentences . . .) analysed in terms of phonology, morphology, and syntax. Our new set of significant cross-linguistic correlations indicates such balancing effects. It suggests trade-offs between facets of phonological complexity and morphological complexity but by no means supports, as could be demonstrated at the end of Section 2.4, the idea of an equal overall complexity in natural languages.

Taking pidgin languages as an example, an attempt was made to conceptualize *semantic complexity* and to relate it to complexity in phonology and morphology. As to the simplicity versus complexity of rigid word order – "rigid word order" in a broader sense – we could at least outline two arguable positions for a necessary future discussion. Both lines of argumentation and all our other empirical and theoretical arguments suggest the view of complexity trade-offs between rather than within the subsystems of language.

References

Altmann, G. 1980. Prolegomena to Menzerath's Law. *Glottometrika* 2: 1–10.

Bett, S. 1999. Can we pin down the number of phonemes in English? *Simpl Speling Newsletter* 1999 (March): 7. (http://victorian.fortunecity.com/ vangogh/555/Spell/phon-inv-art2.html, 12.10.05).

Campbell, G. L. 1991. *Compendium of the World's Languages*. London: Routledge.

Changizi, M.A. 2001. Universal scaling laws for hierarchical complexity in languages, organisms, behaviors and other combinatorial systems. *Journal of Theoretical Biology* 211: 277–295.

Cramer, I.M. 2005. Das Menzerathsche Gesetz. In *Quantitative Linguistik/Quantitative Linguistics*, R. Köhler, G. Altmann & R.G. Piotrowski (eds), 659–688. Berlin: Walter de Gruyter.

Evans, V. 2005. The meaning of *time*: Polysemy, the lexicon and conceptual structure. *Journal of Linguistics* 41: 33–75.

Fenk, A. & Fenk-Oczlon, G. 1993. Menzerath's Law and the constant flow of linguistic information. In *Contributions to Quantitative Linguistics,* R. Köhler & B. Rieger (eds), 11–31. Dordrecht: Kluwer.

Fenk, A. & Fenk-Oczlon, G. 2006. Cross-linguistic computation and a rhythm-based classification of languages. In *From Data and Information Analysis to Knowledge Engineering,* M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger & W. Gaul (eds), 350–357. Berlin: Springer.

Fenk, A., Fenk-Oczlon G. & Fenk L. 2006. Syllable complexity as a function of word complexity. In *The VIII-th International Conference "Cognitive Modeling in Linguistics"*, Vol. 1, V. Solovyev, V. Goldberg & V. Polyakov (eds), 324–333. Kazan: Kazan State University.

Fenk-Oczlon, G. 1983. *Bedeutungseinheiten und sprachliche Segmentierung. Eine sprachvergleichende Untersuchung über kognitive Determinanten der Kernsatzlänge*. Tübingen: Gunther Narr.

Fenk-Oczlon, G. & Fenk, A. 1999. Cognition, quantitative linguistics, and systemic typology. *Linguistic Typology* 3: 151–177.

Fenk-Oczlon, G. & Fenk, A. 2002. The clausal structure of linguistic and pre-linguistic behavior. In *The Evolution of Language out of Pre-Language,* T. Givón & B.F. Malle (eds), 215–229. Amsterdam: John Benjamins.

Fenk-Oczlon, G. and Fenk, A. 2005. Cross-linguistic correlations between size of syllables, number of cases, and adposition order. In *Sprache und Natürlichkeit. Gedenkband für Willi Mayerthaler*, G. Fenk-Oczlon & C. Winkler (eds), 75–86. Tübingen: Narr.

Gell-Mann, M. 1995. What is Complexity? *Complexity* 1: 16–19. (http://www.santafe.edu/~mgm/complexity.html).

Hall, R.A. 1966. *Pidgin and Creole Languages*. Ithaca NY: Cornell University Press.

Heine, B. 1973. *Pidgin-Sprachen im Bantu-Bereich*. Berlin: Dietrich Reimer.

Jespersen, O. 1933. Monosyllabism in English. *Linguistica* In *Selected Writings of Otto Jespersen (no year)*, 574–598. London: George Allen and Unwin.

Juola, P. 1998. Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics* 5 (3): 206–213.

Ladefoged, P. 2001. *Vowels and Consonants: An Introduction to the Sounds of Language*. Oxford: Blackwell.

Lehmann, W. 1978. English: A characteristic SVO language. In *Syntactic Typology*, W. Lehmann (ed.), 169–222. Sussex: The Harvester Press.

McWhorter, J.H. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology* 5(2/3): 125–166.

Maddieson, I. 1984. *Patterns of Sounds*. Cambridge: CUP.

Maddieson, I. 2006. Correlating phonological complexity: Data and validation. *Linguistic Typology* 10: 106–123.

Menzerath, P. 1954. *Die Architektonik des deutschen Wortschatzes*. Hannover: Dümmler.

Meyer, P. 2007. Two semi-mathematical asides on Menzerath-Altmann's law. In *Exact Methods in the Study of Language and Text: dedicated to Professor Gabrid Altmann on the occasion of his 75th birthday*, P. Grzybek & R. Kőhler (eds). New York: Mouton de Gruyter.

Meyer, P. In press. Normic laws in quantitative linguistics. In *The Science of Language. Structures of Frequencies and Relations*, P. Grzybek (ed.).

Mulaik, S.A. 2001. The curve-fitting problem: An objectivist view. *Philosophy of Science* 68: 218–241.

Polikarpov, A.A. 1997. Some factors and regularities of analytic/synthetic development of language system. Paper presented at the XIII International Conference on Historical Linguistics, 10–17 August, Düsseldorf. (http://www.philol.msu.ru/~lex/articles/fact_reg.htm, 13.02.06).

Raukko, J. 2006. Polysemy as complexity? In *A Man of Measure: Festschrift in Honour of Fred Karlsson on his 60th Birthday* [Special Supplement to SKY Journal of Linguistics 19], M. Suominen,

A. Arppe, A. Airola, O. Heinämäki, M. Miestamo, U. Määttä, J. Niemi, K.K. Pitkänen & K. Sinnemäki (eds.), 357–361. Turku: The Linguistic Association of Finland.

Shannon, C.E. 1951. Prediction and entropy in printed English. *Bell System Technology Journal* 30: 50–64.

Simon, H.A. 1996 [1962]. The architecture of complexity: Hierarchic systems. In *The Sciences of the Artificial*. H.A. Simon, 183–216. Cambridge MA: The MIT Press.

Todd, L. & Mühlhäusler, P. 1978. Idiomatic expressions in Cameroon Pidgin English and Tok Pisin. *Papers in Pidgin and Creole Linguistics* 1: 1–35.

Wray, A. & Perkins, M.R. 2000. The functions of formulaic language: An integrated model. *Language & Communication* 20: 1–28.

Zipf, G.K. 1949. *Human behaviour and the principle of least effort. An introduction to human ecology*. Cambridge MA: Addison-Wesley.

Complexity trade-offs in core argument marking

Kaius Sinnemäki
University of Helsinki

Languages have often been claimed to trade off complexity in one area with simplicity in another. The present paper tests this claim with a complexity metric based on the functional load of different coding strategies (head/dependent marking and word order) that interact in core argument marking. Data from a sample of 50 languages showed that the functional use of word order had a statistically significant inverse dependency with the presence of morphological marking, especially with dependent marking. Most other dependencies were far from statistical significance and in fact provide evidence against the trade-off claim, leading to its rejection as a general all-encompassing principle. Overall, languages seem to adhere more strongly to distinctiveness than to economy.

1. Introduction

The recent wake of interest in language complexity has inspired many linguists to develop methods for assessing the complexity of languages (McWhorter 2001; Kusters 2003; Dahl 2004; Hawkins 2004).[1] One of the central interests has been to evaluate the validity of the old claim that all languages are equally complex, expressed in a succinct way by Crystal (1997: 6): "All languages have a complex grammar: there may be relative simplicity in one respect (e.g., no word-endings), but there seems always to be relative complexity in another (e.g., word-position)." In other words, the locus of complexity varies but overall, there is a balance across languages: if one area of grammar (e.g., morphology) is complex, another area (e.g., syntax) is in turn simple, varying from language to language.

1. I would like to thank Fred Karlsson, Matti Miestamo, Seppo Kittilä, and one anonymous reviewer for helpful comments on earlier versions of this paper. Research for this article has been funded by the Academy of Finland under grant 201601 "Complexity in natural languages" and the Finnish Graduate School in Language Studies (Langnet). Their support is gratefully acknowledged. Earlier versions of this paper have been presented at three conferences: ALT VI in Padang, July 21–25, 2005, Approaches to Complexity in Language in Helsinki, August 24–26, 2005, and Leipzig Students' Conference in Linguistics, March 25–26, 2006. Thanks are due to the audiences for helpful comments. I alone am responsible for any remaining errors.

This kind of complexity *trade-off* (also called compensation or balance) is often called for as a proof for equal complexity of languages. Countless scholars have subscribed to similar views without evaluating them more seriously or testing them systematically.

My purpose is to test whether complexity trade-offs are a general all-encompassing principle in language or whether they occur merely in some subdomains or not at all, henceforth called the *trade-off hypothesis*. I suspect trade-offs are more likely between functionally connected variables. For instance, it is intuitively more conceivable that syllable complexity would depend on the size of consonant inventory rather than e.g., the complexity of gender marking. In fact, Maddieson (2006) established a statistically significant positive correlation between the members of the former pair. Moreover, it is unnecessary to study a plethora of features to establish the presence or absence of trade-offs as a general all-encompassing principle. If such generality was to be found, every pair of variables in each area of language should manifest trade-offs, and therefore it should suffice to study even a single area to capture this generality.

To test the hypothesis, I confine myself to studying the interaction of coding strategies involved in the marking of a single functional domain. A functional domain (in the sense of Givón 1981) is a set of closely related semantic or pragmatic functions that are linguistically encoded by at least some languages (e.g., aspect or passive) (cf. Miestamo 2007: 293). Due to this restriction the chosen domain should be as universal as possible. This requirement is met by core argument marking, one of the most "universal" and best documented functional domains in languages. Core arguments are the arguments of a prototypical two-place transitive predicate, one being more agent-like (A) and the other more patient-like (P) (Comrie 2005: 398).

Four linguistic strategies are employed in the marking of core arguments: dependent marking (e.g., case affixes), head marking (e.g., person agreement on the verb), word order, and lexicon.[2] The first three of these are structural strategies and they are the focus of this paper. Although these structural strategies operate on different *additional* functional domains – for example head marking is relevant to referentiality, dependent marking to affectedness (Næss 2004), and word order to focus marking – they interact in core argument marking. According to the trade-off hypothesis, their interaction should not be arbitrary but such that complexity in one strategy is balanced out by simplicity in another. I test this claim statistically with data from 50 languages. The sample is genealogically and areally stratified and randomly chosen; the data is drawn from available reference grammars.

The rest of the paper has the following structure. The language sample is presented in Section 2, followed by a discussion of trade-offs and of the definition of complexity in Section 3. Section 4 formulates the hypothesis for the statistical tests, and in Section 5 I define the parameters to be tested. Section 6 discusses the complexity metric and the data. The results are presented and discussed in Section 7 and conclusions in Section 8.

---

2. Strictly speaking, word order and lexicon do not encode but rather distinguish the core arguments from one another. The role of non-linguistic features, such as context, is not discussed.

Table 1. The distribution and number of genera in the WALS and the language sample.

	Number of genera	
Macro-areas	**WALS**	**Sample**
Africa	64	7
Eurasia	38	4
South East Asia and Oceania	46	5
Australia and New Guinea	125	13
North America	93	10
South America	93	11
Total	459	50

2. Sampling

In order to study trade-offs statistically, a representative and well-balanced random sample is needed. A random and genealogically as well as areally balanced sample is difficult to obtain by sampling mere languages, so it is better to sample groups of languages. For this purpose, Dryer's (1992) method of grouping languages into genera (genealogic subgroupings of languages roughly equivalent with e.g., Germanic or Romance languages) provides a workable solution. He further divides the world into six macro-areas that have relatively equal proportions of genealogical and typological variation. This classification is also used in *The World Atlas of Language Structures* (WALS, Dryer 2005), which I used as a basis for stratification. The total number of genera of spoken languages in the WALS is 459; Table 1 shows their distribution according to macro-areas.[3]

The sample of 50 languages was created by taking from each macro-area a number of genera that is proportional to the total number of genera in that macro-area (Miestamo 2005: 31-39). Each macro-area is therefore represented to the same proportion. The sample is genealogically stratified so that no two languages come from the same genus, subfamily, or family (Table 2).[4] Since creoles have received a lot of attention in the recent discussion on language complexity, one creole was also sampled (one genus was consequently reduced from Australia and New Guinea).

3. The proportions of the sampled genera in North America and South America differ from one another because the one creole sampled, Berbice Dutch Creole, was included in the numbers for South America. Otherwise, the numbers for South America would have been 92 (WALS column) and 10 (sample column).

4. In two macro-areas, in Africa as well as in Australia and New Guinea, small concessions were made at the highest levels of classification due to the unavailability of suitable grammars at the time of sampling.

Table 2. The language sample.

Area	Family	Genus	Language	Source
AFR	Afro-Asiatic	Berber	Middle Atlas Berber	(Penchoen 1973)
		Eastern Cushitic	Somali	(Saeed 1999)
	Khoisan	Central Khoisan	Khoekhoe	(Hagman 1977)
	Niger-Congo	Bantoid	Babungo	(Schaub 1985)
		Southern Atlantic	Kisi	(Childs 1995)
	Nilo-Saharan	Lendu	Ngiti	(Kutsch Lojenga 1994)
		Nubian	Dongolese Nubian	(Armbruster 1960)
EUR	Dravidian	Southern Dravidian	Kannada	(Sridhar 1990)
	Finno-Ugric	Ugric	Hungarian	(Rounds 2001)
	Indo-European	Celtic	Welsh	(King 1993)
	Kartvelian	Kartvelian	Georgian	(Harris 1981)
ANG	Australian	Bunuban	Gooniyandi	(McGregor 1990)
		Maran	Warndarang	(Heath 1980)
		Pama-Nyungan	Diyari	(Austin 1981)
	Dagan	Dagan	Daga	(Murane 1974)
	Kuot	Kuot	Kuot	(Lindström 2002)
	Sepik-Ramu	Lower Sepik	Yimas	(Foley 1991)
	Sko	Western Sko	Sko	(Donohue 2004)
	Solomons East Papuan	Solomons East Papuan	Lavukaleve	(Terrill 2003)
	Torricelli	Kombio-Arapesh	Arapesh	(Conrad and Wogiga 1991)
	Trans-New Guinea	Awju-Dumut	Kombai	(de Vries 1993)
		Timor-Alor-Pantar	Adang	(Haan 2001)
	West Papuan	North-Central Bird's Head	Maybrat	(Dol 1999)
	Yele	Yele	Yelî Dnye	(Henderson 1995)
SAO	Austro-Asiatic	Munda	Korku	(Nagaraja 1999)
	Austronesian	Sundic	Indonesian	(Sneddon 1996)
	Hmong-Mien	Hmong-Mien	Mien	(Court 1985)
	Sino-Tibetan	Qiangic	Qiang	(LaPolla 2003)
	Tai-Kadai	Kam-Tai	Thai	(Iwasaki and Ingkaphirom 2005)
NAM	Algonquian	Algonquian	Plains Cree	(Dahlstrom 1991)
	Eskimo-Aleut	Eskimo-Aleut	West Greenlandic	(Fortescue 1984)
	Hokan	Yuman	Maricopa	(Gordon 1986)

(*Continued*)

Table 2. Continued.

Area	Family	Genus	Language	Source
	Mayan	Mayan	Tzutujil	(Dayley 1985)
	Muskogean	Muskogean	Koasati	(Kimball 1991)
	Na-Dene	Athapaskan	Slave	(Rice 1989)
	Penutian	Miwok	Southern Sierra Miwok	(Broadbent 1964)
	Siouan	Siouan	Osage	(Quintero 2005)
	Uto-Aztecan	Corachol	Cora	(Casad 1984)
	Wakashan	Southern Wakashan	Nuuchahnulth	(Nakayama 2001)
SAM	Aymaran	Aymaran	Jaqaru	(Hardman 2000)
	Cariban	Cariban	Hixkaryana	(Derbyshire 1979)
	Chibchan	Aruak	Ika	(Frank 1990)
	Mura	Mura	Pirahã	(Everett 1986)
	Panoan	Panoan	Shipibo-Konibo	(Valenzuela 1997)
	Peba-Yaguan	Peba-Yaguan	Yagua	(Payne and Payne 1990)
	Quechuan	Quechuan	Imbabura Quechua	(Cole 1985)
	Trumai	Trumai	Trumai	(Guirardello 1999)
	Tupian	Tupi-Guaraní	Urubú-Kaapor	(Kakumasu 1986)
	Warao	Warao	Warao	(Romero-Figueroa 1997)
	Creole	Creole	Berbice Dutch Creole	(Kouwenberg 1994)

3. The definition of complexity

In order to define and measure complexity in a meaningful way, a good understanding of the notion of trade-off is first needed. Whereas trade-offs may be viewed simply as negative correlations between the complexities of different grammatical structures (however measured), it is more instructive to consider the underlying principles that could trigger them. Two underlying principles for trade-offs may be recognized, namely economy and distinctiveness. Both principles relate to the well-established principle of one-meaning–one-form but from different perspectives.

The principle of economy may be simply defined as one meaning encoded by only one form. Encoding via more than one form is avoided as well as unnecessary marking of distinctions. Violations of economy increase synonymy and redundancy (and hence complexity, Dahl 2004: 9–11, 51–55, 120–121) (see Miestamo, this volume, for detailed sub-criteria of the violations of the one-meaning–one-form principle). If in language change a new form emerges to compete for the same niche with an older one, either should be dispensed with: free variation between the strategies and the partitioning of the functional load violate the principle of economy (cf. Dahl 2004: 128–134). The

principle of distinctiveness, on the other hand, is defined as one meaning encoded by at least some form. Distinctiveness is violated by one meaning being encoded by a non-unique form, i.e., if meaning distinctions made in the grammar are encoded by identical forms. Such insufficient encoding of meaning distinctions leads to homophony and ambiguity and increase processing difficulties (inefficiency in terms of Hawkins 2004). If in language change a coding strategy is lost, a new strategy should emerge to take over so that meaning distinctions are not lost.

The notion of trade-off assumes that the interaction between coding strategies interacting in the same functional domain is equally controlled by economy and distinctiveness: their combined outcome gears grammar towards adhering to the principle of one-meaning–one-form. Although their operation may be interpreted from the point of view of speaker and hearer preferences, I follow Dahl (2004: 39) in keeping difficulty and complexity apart from one another: difficulty is a matter of language use while complexity is an absolute property of the language system.

Complexity is typically measured as the number of parts or rules in a system (cf. Miestamo, this volume, for a detailed discussion). Less often the metrics have paid any attention to the functional load of transparent marking. Functional load is the measure of how often a contrast – a particular meaning distinction made with particular forms – is employed in a language. For instance, Hungarian always marks P with case and thus has a greater functional load than e.g., English, which has case marking of P only for personal pronouns. From an information theoretic point of view, great functional load means high frequency and thus great expectedness – a sign of low complexity. However, a metric based on such a definition of complexity would yield English case marking of P more complex than that of Hungarian. If, on the other hand, we take into account the asymmetries in the frequency and structural marking of the arguments, we come closer to a complexity metric that pays attention to functional loads.

In core argument marking, A is most frequently animate and definite whereas P is inanimate and indefinite and both typically lack overt structural marking. A is less frequently inanimate and indefinite and P animate and definite but these are more often structurally marked (Comrie 1989). In most contexts, A and P could be distinguished solely by semantic properties: where semantic properties satisfy distinctiveness, structural marking is unnecessary according to economy. According to Haspelmath (2006) however, structural asymmetries may be explained via frequency asymmetries instead of relying on (semantic) markedness: the most frequent members of a category are typically not structurally marked (probably due to their high expectedness), whereas less frequent members are. Arguments are thus expected to be distinguished only when A is inanimate and/or indefinite or P is animate and/or definite. Consequently, marking that occurs with the most frequent A and P is redundant and thus more complex (but in some ways also rather unexpected). Thus, the greater the functional load of a coding strategy, the more redundant and therefore the more complex it is.

For the present purposes, it is most relevant to measure how often a particular coding strategy distinguishes A from P. This translates roughly into a measure of differential

argument marking. The details of measuring the functional load of the coding strategies will be discussed in Section 6.

4. Hypothesis for the statistical tests

Having discussed the nature of trade-offs, we may now formulate the hypothesis for statistical tests:

> the complexity values of different coding strategies depend on one another in an inverse way.

According to this formulation, the complexity values of different coding strategies should correlate negatively and in a statistically significant way. If a correlation is not statistically significant, the sign of correlation (+/–) might still be relevant. If the correlation is positive, we must reject the hypothesis. The closer the correlation is to zero, the smaller the relationship is between the variables leading to the rejection of the hypothesis.

Things are not this simple, however. Research by Fenk-Oczlon and Fenk (this volume) has indicated that trade-offs are more likely between coding strategies from different coding domains (e.g., one morphological and the other syntactic) than between strategies form the same coding domain. The hypothesis should be slightly modified to accommodate these tendencies:

> a negative correlation is expected for a pair of variables from different coding domains but a positive correlation is possible only domain-internally.

5. Coding strategies: Definitions and constraints

Three constraints, typical in word order studies, are observed in the analysis of the coding strategies. Firstly, the focus is on simple active affirmative indicative main clauses; marking outside this focus is excluded. Secondly, the marking involved with pronouns is excluded as well, ruling out coding strategies which occur only with pronouns or in the absence of full NPs. Although clauses with two full core NPs are rare in the spoken form of most languages (Du Bois 1987), two reasons justify excluding pronouns from the analysis. For one, this variation in argument marking is a consequence of how languages mark the functional domain of topic continuity; focusing on core argument marking of full NPs rules out the effects of a neighbouring functional domain that interacts with it. A more serious reason is that especially in the case of clitics it may be impossible to determine whether a pronoun instantiates head or dependent marking. A third constraint is that an argument is understood as being part of the clause proper only if it is not obligatorily separated from the rest of the clause by a pause and if there is no pronoun in situ replacing a transposed argument. Although earlier studies have

noted that word order rules increase complexity, they have not been subjected to systematic cross-linguistic measures before.

The three structural coding strategies under investigation have traditionally been labelled case marking, agreement, and word order (WO). Since the first two terms have typically been associated with affixal morphology, I adopt and slightly modify Nichols' (1992) terms head and dependent marking, which cover both affixal and non-affixal marking.

For the present purpose, I define dependent marking (DM) and head marking (HM) in the following way. An argument NP is dependent-marked when an NP functions as a core argument (A or P) of a prototypical two-place transitive predicate and occurs in a particular form distinct from that of the other argument. The definition measures the degree of differential argument marking and therefore the arguments are treated together also in the statistical tests, i.e., only one column stands to represent dependent marking in Table 3. The definition pays no attention to the formal way in which the argument is marked as dependent. For instance, Kannada (1) uses an accusative suffix *-annu*, Urubú-Kaapor (2) uses a postposition *ke*, but Somali (3) uses different tones (*Cáli* in citation form, with high-low tone pattern, and *Cali* in the nominative, with low-low tone pattern).

Kannada (Southern Dravidian, Sridhar 1990: 86)

(1) *cirate* *mariy-annu* *nekkutti-de*
Leopard cub-ACC lick.PRS-3SG.N
'The leopard is licking the cub.'

Urubú-Kaapor (Tupi-Guaraní, Kakumasu 1986: 351)

(2) *pe* *tapĩ* *xaʾe* ***ke*** *kutuk* *tī*
and Tapĩ Xaʾe OBJ 3.pierce also
'And Tapĩ also pierced Xaʾe (with an arrow).'

Somali (Eastern Cushitic, Saeed 1999: 59, 65, 229)

(3) ***Cali*** *warqáddíi* ***wuu*** *íi* *dhiibay*
Ali.NOM letter.ABS DECL.3 1SG.to pass.PST
'Ali passed the letter to me.'

Sometimes the formal marker occurs on some other constituent than the one whose role it indicates. In Yagua (4), the clitic *-níí* marks P as third person singular. However, it is attached neither to the predicate nor the NP itself but to the constituent preceding the NP. Since P functions as a reference point for the clitic placement, I treat it as dependent-marked.

Yagua (Peba-Yaguan, Payne and Payne 1990: 255)

(4) a. ***sa****-jimyiy* *Alchíco-****níí*** *quiiva*
3SG.A-eat Alchico-3SG.P fish
'Alchico is eating the fish.'

b. ***sa****-jimyiy-****níí*** *quiiva*
3SG.A-eat-3SG.P fish
'He is eating the fish.'

Head marking is defined as follows: an argument NP of a prototypical two-place transitive predicate is head-marked when the form of the predicate transparently determines that the argument has (a) certain feature(s). According to this definition the form of the predicate does not determine the role of the argument, but rather its form determines that the argument has particular properties: e.g., the A argument has properties third person singular neuter in (1). The definition pays no attention to the formal ways in which the features of the argument are marked on the head. For instance, Kannada (1) uses the suffix *-de*, Yagua (4) uses the proclitic *sa-*, but in Somali (3) the marking does not occur on the main verb, but on the declarative marker *waa*, which has fused with the third person singular marker to yield *wuu*. In the minimal case, the absence of overt formal marking (i.e., zero marking) on the predicate often indicates that the argument is in the third person. Languages also vary in how many features of the argument the predicate determines. Some determine only person (Urubú-Kaapor (2)), others person and number (Somali (3)), and still others person, number, and gender (Kannada (1)).

The third morphosyntactic strategy is word order. It distinguishes the core arguments when the role of the noun argument is determined by its position relative to the predicate, whereby the arguments may not occupy each other's position. In other words, if any of the *reversible word order pairs* AVP/PVA, APV/PAV, and VAP/VPA occur in a language, word order cannot distinguish the roles of the arguments. For instance, Thai (5) employs AVP order when the A is topical: reversing the order of the NPs changes their roles and thus PVA is impossible.

Thai (Kam-Tai, Iwasaki and Ingkaphirom 2005: 110)

(5) *lék tè nɔ̀ɔy*
Lek kick Noy
'Lek kicks Noy.' *'Noy kicks Lek.'

Thai may also topicalize the P by placing it clause-initially, but since this does not lead to a reversible order – AVP and PAV do not form a reversible word-order pair – word order distinguishes core arguments from one another. One should, of course, check all possible reversible word order pairs, not only the reversible word order pair involving the canonical word order. In Koasati, the canonical word order is APV but its reverse (PAV) does not occur (Kimball 1991: 513–517). Because another reversible word order pair is allowed, namely AVP and PVA, it is unlikely that word order distinguishes the roles of the arguments in Koasati at all.

Some languages allow for a reversible word order pair but either one or both orders are marked by changing the properties of head/dependent marking that occur in the canonical word order. By changes I mean removal of overt marking or a new set of markers, i.e., some marking which does not occur in the canonical order. Since the change in morphological marking parallels a change in word order, these orders would probably not be allowed for without the change in morphological marking. This leads us to think that word order helps to distinguish the roles in the canonical order. As an example, consider word order in Slave (6).

Slave (Athapaskan, Rice 1989: 1197)

(6) a. *lı̨ ʔehkee kayı̨hshu*
dog (A) boy (P) 3.bit
'The dog bit the boy.'

b. *ʔehkee lı̨ kayeyı̨hshu*
boy (P) dog (A) 3.bit.4
'The boy, a dog bit him.'

The canonical order is APV (6a), but when the object NP occurs clause-initially (6b), i.e., in its non-canonical position, the predicate is head-marked with the co-indexing pronoun *ye-*.

The three strategies defined in this section are morphosyntactic strategies. However, they are not the only *linguistic* strategies: the inherent lexical features of the arguments are often used as well. Some languages employ morphosyntactic marking only when the lexical features fail (Kittilä 2005). For instance, the agent is optionally dependent-marked in Qiang only when P is moved to a clause initial position or higher on the person hierarchy than A (7).

Qiang (Qiangic, LaPolla 2003: 78, 79, 122)

(7) a. *qa ʐawa ɦo-ɣlu-a*
1SG stone DIR-roll-1SG
'I rolled the stone down.'

b. *mi-wu qa zə-dʑi*
person-A 1SG DIR-hit
'Somebody hit me.'

c. *the:-tɕ pi:-xsə-la sum-wu de-l-ji ŋuə*
3SG-GEN pen-three-CLF teacher-A DIR-give-CSM COP
'*The teacher* gave him three pens.'

This paper focuses on morphosyntactic encoding and excludes the role of the lexicon in core argument marking.

6. Measuring functional load – and complexity

In Section 2, complexity of a coding strategy was defined as its functional load (FL). FL could of course be counted from actual texts but this was not possible within the limits of this paper. I rather estimated the FL of a coding strategy based on the grammar descriptions of each language.

Functional load forms a continuum whose end-poles are defined as *none* – a coding strategy never distinguishes the arguments in a particular language – and *full* – a coding strategy always distinguishes the arguments from one another. The FL of coding strategies is limited in various ways whereby it would be uninformative to assign

the same values to all coding strategies with limited FL. The FL of argument marking is also often split between the coding strategies, even in a complementary fashion. Consequently, the complexity metric should capture trade-offs regardless of the FL being carried by a single strategy or by more than one complementary strategy. This is achieved by distinguishing *marginal* FL from *extensive* FL; impressionistically, the FL of a coding strategy is marginal when it distinguishes the arguments in less than half of the contexts, while it is extensive when it encodes them in roughly half or more of the contexts.

The values marginal and extensive are more precisely defined in terms of typological markedness (Croft 1990). Typological markedness is defined using three criteria one of which is the frequency criterion: the unmarked member of a category occurs at least as frequently as the marked member in a given language. Using the frequency criterion helps us to avoid problems with semantically based definitions of both markedness (Haspelmath 2006) and differential argument marking (cf. Næss 2004). The marked member never occurs more frequently than the unmarked member. Translating this into our metric, e.g., the most frequent P (inanimate and indefinite) is unmarked and it never occurs less frequently than the marked P (animate and definite). If a coding strategy encodes an argument when it is expected to, that is, when it is marked (and the least frequent), it is analysed as having marginal FL. If a coding strategy encodes an unmarked argument, it has extensive FL. For instance, in Kannada human patients are dependent-marked but because non-humans are optionally dependent-marked as well (Sridhar 1990: 86), the functional load of dependent marking is analysed as extensive. In (6) we saw that Slave uses word order to encode the core arguments only in the canonical (i.e., the most frequent and thus unmarked) order. Consequently, the FL of word order in Slave is extensive.

When the frequency criterion of typological markedness is irrelevant to the type of restriction, FL will be estimated by counting the number of relevant language specific contexts and estimating whether e.g., optionality refers to the evoking or the dropping of marking. For instance, Arapesh marks the person, number, and class of the patient on the head in transitive verb classes 1 and 2 but optionally in classes 3 and 4.[5] Class 1 is relatively small but classes 2–4 are rather large: this would indicate a marginal FL. However, Conrad and Wogiga (1991: 78) note that ". . . [w]ith relatively few exceptions, verb morphology is the same for all verbs, regardless of their distribution in a sentence and regardless of different patterns of affixation [of the verb classes, KS] . . . " Dropping the verb morphology seems rather an exception than the norm in Arapesh, which leads us to analyse the FL of patient head-marking as extensive.

Note that the metric is able to capture complementary functional loads. In West Greenlandic the arguments are analysed as dependent-marked when the number of

5. Also in verb classes 7 and 8, but these classes contain only one verb each.

either or both is singular but not when both arguments are plural because the case marker for plural agent and patient is homophonous *-(i)t* (Fortescue 1984: 206–210). This results in extensive FL for dependent marking.[6] The noun arguments cannot occur in reversed order in this situation but only in AP order. The restrictions on the FL of word order are the mirror image of the restrictions of dependent marking, resulting in marginal FL.

The instances of marginal and extensive FL are described in detail in the lists below, containing only the ones not already discussed.

Marginal FL:

1. Adang: the person of P is head-marked by a small closed class of rather frequently used verbs. The arguments are also dependent-marked when focused, A optionally but P obligatorily.
2. Ika (A), Shipibo-Konibo (A), and Slave (P): plural number of the (parenthesized) argument is optionally head-marked.
3. Kannada: WO distinguishes the arguments when both are inanimate.
4. Korku: the person and number of animate patients are head-marked.
5. Sko: the gender and number of the patient are optionally head-marked when the patient is possessed.
6. Slave: A is head-marked when third person singular and in perfective mode but only with one conjugation marker, and optionally as third person plural when animate.
7. Urubú-Kaapor: P is optionally dependent-marked.
8. Yagua: (roughly) definite patients are dependent-marked.

Extensive FL:

9. Cora, Kuot, Slave, Trumai, Yagua: word order distinguishes the arguments in canonical order.
10. Gooniyandi: inanimate agents are dependent-marked, animate agents optionally. The person and number of the patient are head-marked when A is plural and when one classifier is in future and present tenses.
11. Ika (P), Qiang, and Sko: when arguments could be confused (e.g., non-canonical WO), either is dependent-marked, optionally elsewhere.
12. Kisi: A is dependent-marked with a subject noun class pronoun, optionally for animate singular. P is optionally dependent-marked with object noun class pronouns. There is also tonal downstep between the verb and the following direct object.

6. Note that our definition of functional load is hearer-oriented: homophonous forms violate the principle of distinctiveness.

13. Korku: animate patients are dependent-marked, inanimate optionally.
14. Lavukaleve: the person and number of the agent are head-marked in present tense and in the habitual aspect, optionally elsewhere.
15. Somali: a non-focused P is dependent-marked with a tone pattern distinct from that of A.
16. Warao: the person and number of the agent are head-marked when aspect is not marked or when tense is either implied or marked.
17. Warndarang: the person (and number) of the patient is head-marked when the number of either or both arguments is plural.
18. Yelî Dnye: the person (and partially number) of the agent is head-marked for continuous aspect and partially for punctiliar aspect.
19. Yimas: the person and number of the arguments are obligatorily head-marked when they are new referents but optionally when established.

Two problematic cases not treated in the lists are discussed here. In Daga, an animate P is head-marked when third person singular but inanimate patients are only optionally marked as totally affected (Murane 1974: 44). The problem is how to assess the FL of such a heterogeneous marking. Contrasting this FL with one that has already been analysed might help. Obligatory head marking in Daga is more limited than e.g., obligatory dependent marking in Kannada. Optional head marking in Daga, on the other hand, is impressionistically rather similar in scope with optional dependent marking in Kannada. Since FL of dependent marking in Kannada was analysed as extensive, FL of patient head-marking cannot be analysed as extensive but rather as marginal. In Maricopa the number of the patient is head-marked by suppletive stems for some verbs and by an optional plural prefix (Gordon 1986: 22, 100). Since the source did not specify what type of verbs had suppletive verb forms for number, I interpreted "some" literally and classified FL as marginal.

For the purposes of the statistical tests, the four FL values – none, marginal, extensive, full – were converted into ordinal values 1, 2, 3, and 4, respectively (Table 3).

Table 3. The data.

Language	WO_A/P	DM_A/P	HM_A	HM_P
Adang	none	marg	none	marg
Arapesh	full	none	full	ext
Babungo	none	none	none	none
Berbice Dutch Creole	full	none	none	none
Cora	ext	none	full	full
Daga	none	none	full	marg
Diyari	none	full	none	none
Georgian	none	full	full	none
Gooniyandi	none	ext	full	ext

(*Continued*)

Table 3. Continued.

Language	WO_A/P	DM_A/P	HM_A	HM_P
Hixkaryana	full	none	full	full
Hungarian	none	full	full	full
Ika	none	ext	marg	full
Indonesian	full	none	full	none
Jaqaru	none	full	full	full
Kannada	marg	ext	full	none
Khoekhoe	none	full	none	none
Kisi	full	ext	none	none
Koasati	none	full	full	none
Kombai	full	none	full	none
Korku	none	ext	none	marg
Kuot	ext	none	full	full
Lavukaleve	full	none	ext	full
Maricopa	none	full	full	marg
Maybrat	full	none	full	none
Middle Atlas Berber	full	full	full	none
Mien	full	none	none	none
Ngiti	full	none	full	none
Dongolese Nubian	none	full	full	none
Nuuchahnulth	none	none	full	none
Osage	full	full	full	full
Pirahã	full	none	none	none
Plains Cree	none	none	full	full
Qiang	none	ext	full	full
Quechua	none	full	full	none
Shipibo-Konibo	none	full	marg	none
Sko	full	ext	full	marg
Slave	ext	none	marg	marg
Somali	none	ext	full	none
Southern Sierra Miwok	none	full	full	none
Thai	full	none	none	none
Trumai	ext	full	none	none
Tzutujil	none	none	full	full
Urubú-Kaapor	none	marg	full	none
Warao	none	none	ext	none
Warndarang	full	none	full	ext
Welsh	full	full	full	none
West Greenlandic	marg	ext	full	full
Yagua	ext	marg	full	none
Yelî Dnye	full	full	ext	full
Yimas	none	none	ext	ext

7. Results and discussion

According to the hypothesis, statistically significant trade-offs should exist between the different variables involved in core argument marking. Word order (WO), dependent marking (DM) and the head-marking of agent (HM_A) and patient (HM_P) were taken as variables as such. Other variables were the "normalized" head marking (HM_N) and morphological marking (the sum of dependent marking (DM) and head marking (HM_N)). These five variables combine into nine pairs (Table 4). HM_N means that the functional load of head marking is normalized to the lowest value of head marking of A/P in each language, roughly measuring the degree of differential argument marking by HM (or "rich agreement"). For instance, the FL of HM_A e.g., in Maricopa is 3 and that of HM_P is 1, but the FL of normalized HM is 1.

To test the hypothesis, I used Kendall's (tau) nonparametric correlation test in the open-source statistical computing environment R (R Development Core Team 2006). Kendall's tau measures the difference between the probabilities that two variables are in the same or different order. If Kendall's tau was statistically significant for a particular pair of variables, the Chi-Square test and its tabulated data was used to double-check the result and help in the interpretation. For the Chi-Square test, the values of the coding strategies were classified into two groups: a) those with no FL and b) those with non-zero FL, corresponding straightforwardly to "simple" and "complex", respectively.

Table 4. The Kendall's tau values for each pair of variables.

Pair of variables	tau	p
WO–DM	–0.302	0.017
WO–HM_A	–0.035	0.787
WO–HM_P	–0.038	0.768
WO–HM_N	–0.005	0.967
WO–M	–0.249	0.041
DM–HM_A	0.066	0.606
DM–HM_P	–0.094	0.452
DM–HM_N	–0.114	0.365
HM_A–HM_P	0.201	0.118

Table 4 summarizes the results for Kendall's tau. The values were statistically significant only for two correlation pairs: those between WO and morphological marking and WO and DM. The correlation between head marking of A and P was strong enough to validate closer inspection. For all the other pairs Kendall's tau was less than ± 0.12 and far from being statistically significant. In fact some values (especially

that between WO and HM_N) were so close to zero as to render the trade-off hypothesis vacuous: almost no interaction occurred between these variables. We may still observe that the sign of correlation conformed to the hypothesis for all non-significant pairs: WO had a very small negative correlation with HM_A, HM_P, and HM_N, whereas DM correlated positively with head marking of A but negatively with HM_ P and HM_N.

The variables representing the different coding domains (WO and morphological marking) correlated negatively in a statistically significant way (tau = –0.249, $p < 0.05$). The Chi-Square value was not statistically significant but still close to it ($p = 0.088$). According to Table 5, eight sample languages relied exclusively on word order and three lacked full FL for any coding strategy. Two of these, Warao and Nuuchahnulth, had extensive and full head marking of A, respectively, but none for P. Babungo, on the contrary, had no head or dependent marking (a floating tone might serve as a dependent marker, but there was not enough data to confirm this). It also allows for a reversible word order pair AVP/PVA: the canonical word order is AVP, but the reverse PVA may be used to topicalize the P. Thus, only context tells when the sentence is AVP or PVA (8):

Babungo (Bantoid, Schaub 1985: 141, 250 n1)

(8) *Wèe wî jìa Làmbí*
child that hold-PFV Lambi
'As for that child, Lambi held him.'
or 'That child held Lambi.'

It seems almost a norm for languages to have at least some morphological marking of the core arguments. The absence of morphological marking seems a good indicator that WO has some FL but the presence of morphological marking indicates only weakly that WO has a zero FL. The overall conclusion is that the smaller the FL of morphological marking is the more likely WO carries a greater functional load in core argument marking. This result conforms well to the trade-off hypothesis.

The correlation between WO and DM was –0.302 ($p < 0.05$). The Chi-Square value was also statistically significant ($p < 0.05$). Interpreting from Table 5, the absence of DM seems a good indicator of the presence of WO, and the absence of WO seems a good indicator of the presence of DM. It is probably safe to conclude that the smaller the functional load of DM, the more likely WO has a greater functional load, and the greater the functional load of DM, the less likely WO has a smaller (rather than greater) functional load. A strong inverse dependency occurs between these variables. It is probably no coincidence they have often been seen as complementary devices in argument marking (e.g., Vennemann 1973). Since the correlation of WO and HM was so small, the correlation between WO and DM explains why WO correlated also with overall morphological marking.

Table 5. Chi-square values for a few pairs of variables.

	M–	**M+**	**DM–**	**DM+**
WO–	3	22	7	18
WO+	8	17	15	10
	χ^2 =2.91, p = 0.088		χ^2 = 5.19, p = 0.023	
	HM_A–	**HM_A+**	**WO/DM–**	**WO/DM+**
HM_P–	9	18	5	22
HM_P+	2	21	12	11
	χ^2 = 4.39, p = 0.036		χ^2 = 6.27, p = 0.012	

Since the data contained many tied values, the reliability of Kendall's tau was further studied with a resampling procedure called bootstrap.[7] This procedure was performed for the statistically significant pairs WO and DM and WO and morphological marking. Languages were randomly sampled with replacement 10,000 times and the value of Kendall's tau was computed each time. Confidence limit for 0.95 was computed by extracting the 250th and 9750th resampled tau values. The confidence limits at level 0.95 were [–0.249 and 0.250] for WO and DM and [–0.244 and 0.246] for WO and morphological marking thus confirming the confidences of the original values (–0.302 and –0.249).

Although the correlation between the head-marking of A and P (tau = 0.201) was not statistically significant, the Chi-Square value was ($p < 0.05$). The presence of HM_P seems a good indicator of the presence of HM_A, and the absence of HM_A seems a good indicator of the absence of HM_P. This positive dependency between the variables provides evidence for dependencies between the head-marking strategies. The result reflects the agreement hierarchy (subject < object), which tells that the presence of object agreement implies the presence of subject agreement but subject agreement can exist without object agreement (Croft 1990: 101–107)

The trade-off hypothesis predicts that HM should correlate strongly with both WO and DM. Notably neither variable had a significant relationship with the head-marking of A or P or their neutralized sum. The strong dependency between WO and DM might suggest languages prefer distinguishing the core arguments via either of them and rely on HM only when they are insufficient. Accordingly, the trade-off hypothesis was adjusted to expect HM when core arguments were not fully distinguished by WO and DM. Since HM_A is so frequent cross-linguistically (occurred in 39 sample languages), HM_P was expected instead (note that it often implies the presence of HM_A as well). This adjusted prediction was borne out by the results (Table 5): when WO and

7. I am grateful to Stefan Werner for advice with the bootstrap method.

DM fully distinguish the core arguments (WO/DM+, i.e., their combined FL is 3 or more), HM_P is less likely to occur than not, and when WO and DM cannot fully distinguish the core arguments (WO/DM-) HM_P is more likely to occur than not. The Chi-Square value ($\chi^2 = 6.27$) for this inverse dependency was statistically significant ($p < 0.05$). HM_A seems to occur in languages regardless of its "dispensability" in core argument marking, but HM_P does so less often.

Finally, a few words should be said about how languages overall conformed to the principles of economy and distinctiveness. The functional loads of the coding strategies (WO, DM, and HM_N) were added up in each language. According to the principle of distinctiveness, languages should distinguish core arguments in all contexts. In other words, the sum of the FLs should not be less than 3. In nine sample languages the overall FL was less than this, violating the principle of distinctiveness. The most notable exceptions were Warao, Nuuchahnulth, and Babungo (already discussed above). In 22 languages, the overall functional load was exactly 3, adhering perfectly to the two principles. The remaining 19 languages (38 %) violated the principle of economy to varying degrees, most notably Osage and Yelî Dnye, which exhibit at least extensive FL for all coding strategies.[8] Overall, adherence to distinctiveness seems stronger than adherence to economy, which is another way of saying that languages allow for redundancy more than ambiguity.

8. Conclusions

The claim that languages trade off complexity in one area with simplicity in another has been criticised in the recent discussion on language complexity (McWhorter 2001; Kusters 2003; Shosted 2006). This study is one of the first attempts at investigating the claim with morphosyntactic variables and with a large balanced sample. The results justify rejecting trade-offs as an all-encompassing principle in languages: most of the correlations were small or even approaching zero, indicating no relationship between the variables.

However, the results support the trade-off hypothesis in a weaker form in that trade-offs occur in some subdomains: word order correlated in a statistically significant way with morphological marking and dependent marking. The statistically significant positive dependency between the head-marking of agent and patient further conforms to the predictions of the hypothesis. So does the sign of correlation (+/–), which was as predicted for all pairs of variables. Lastly, the presence of the head-marking of P depended in an inverse and statistically significant way on whether word order and dependent marking fully distinguished the core arguments. Overall, it would be fair to

8. I am a bit sceptical about the analysis of Yelî Dnye and Osage, since the sources did not cover word order phenomena in great detail.

say that languages manifest at least some trade-offs in core argument marking, most notably that between WO and DM.

The results agree with the work of Fenk-Oczlon and Fenk (this volume, and the references therein). They have established several negative correlations between e.g., the "complexities" of syllable structure, word structure, and clause structure counted in terms of numbers of phonemes, syllables and words. The results disagree with Shosted (2006), who found no evidence for trade-offs in a cross-linguistic study on the complexities between syllable structure and inflectional synthesis on the verb. However, this discrepancy in our results might stem from the choice of parameters: I intentionally chose variables that were functionally connected whereas the parameters studied by Shosted (2006) were functionally rather dissimilar. Trade-offs seem limited to functionally connected variables but even then they are not an all-encompassing principle.

It is not at all clear that equal complexity of languages could be shown by a synchronic snapshot on trade-offs alone. In addition to trade-offs, one should show for instance that simplifying and complexifying tendencies are equally frequent in language change. In other words, languages should follow the principles of economy and distinctiveness to the same degree both synchronically and diachronically. What the present study on trade-offs has hopefully shown is to what degree languages adhere to the principles of economy and distinctiveness synchronically in marking the core arguments. According to the results, most languages conform to the principle of one-meaning–one-form, but those that do not conform to it adhere to distinctiveness at the expense of economy. This may reflect the tendency in languages to accumulate complexity during their long histories, i.e., that they exhibit fewer simplifying than complexifying tendencies (McWhorter 2001; Dahl 2004), but it may as well reflect the tendency in small, isolated, and tight-knit communities to preserve complexity, i.e. that they exhibit fewer simplifying tendencies than languages spoken by large communities with loose social networks and adult language learning by outsiders (Kusters 2003: 5–9).

Abbreviations

1	first person
3	third person
4	fourth person (only in Slave)
A	agent
ABS	absolutive
ACC	accusative
CLF	classifier
COP	copula
CSM	change of state marker
DECL	declarative
DIR	directional prefix

GEN	genitive
N	neuter
NOM	nominative
OBJ	object marker
P	patient
PST	past tense
PFV	perfective
PRS	present tense
SG	singular

References

Armbruster, C.H. 1960. *Dongolese Nubian: A Grammar*. Cambridge: CUP.

Austin, P. 1981. *A Grammar of Diyari, South Australia*. Cambridge: CUP.

Broadbent, S.M. 1964. *The Southern Sierra Miwok Language*. Berkeley CA: University of California Press.

Casad, E.H. 1984. Cora. In *Studies in Uto-Aztecan Grammar*, Vol. 4: *Southern Uto-Aztecan Grammatical Sketches*, R.W. Langacker (ed.), 153–459. Dallas TX: Summer Institute of Linguistics.

Childs, G.T. 1995. *A Grammar of Kisi: A Southern Atlantic Language*. Berlin: Mouton de Gruyter.

Cole, P. 1985. *Imbabura Quechua*. London: Croom Helm.

Comrie, B. 1989. *Language Universals and Linguistic Typology*. 2nd edn. Oxford: Blackwell.

Comrie, B. 2005. Alignment of case marking. In *The World Atlas of Language Structures*, M. Haspelmath, M. Dryer, D. Gil & B. Comrie (eds), 398–405. Oxford: OUP.

Conrad, R.J. & Wogiga, K. 1991. *An Outline of Bukiyip grammar*. Canberra: Australian National University.

Court, C. 1985. *Fundamentals of Iu Mien (Yao) Grammar*. PhD Dissertation, University of California at Berkeley.

Croft, W. 1990. *Typology and Universals*. Cambridge: CUP.

Crystal, D. 1997. *The Cambridge Encyclopedia of Language*. 2nd edn. Cambridge: CUP.

Dahl, Ö. 2004. *The Growth and Maintenance of Linguistic Complexity*. Amsterdam: John Benjamins.

Dahlstrom, A. 1991. *Plains Cree Morphosyntax*. New York NY: Garland Publishing.

Dayley, J. 1985. *Tzutujil Grammar*. Berkeley CA: University of California Press.

Derbyshire, D.C. 1979. *Hixkaryana*. Amsterdam: North-Holland.

Dol, P. 1999. *A Grammar of Maybrat: A Language of the Bird's Head, Irian Jaya*. PhD Dissertation, University of Leiden.

Donohue, M. 2004. Ms. A Grammar of the Skou Language of New Guinea. National University of Singapore. [http://www.papuaweb. org/dlib/tema/bahasa/skou/ (November 13, 2006)].

Dryer, M. 1992. The Greenbergian word order correlations. *Language* 68: 81–138.

Dryer, M. 2005. Genealogical language list. In *The World Atlas of Language Structures*, M. Haspelmath, M. Dryer, D. Gil & B. Comrie (eds), 584–644. Oxford: OUP.

Du Bois, J. 1987. The discourse basis of ergativity. *Language* 63: 805–855.

Everett, D.L. 1986. Pirahã. In *Handbook of Amazonian Languages,* Vol. 1, D.C. Derbyshire & G.K. Pullum (eds), 200–325. Berlin: Mouton de Gruyter.
Foley, W.A. 1991. *The Yimas Language of New Guinea*. Stanford CA: Stanford University Press.
Fortescue, M. 1984. *West Greenlandic*. London: Croom Helm.
Frank, P. 1990. *Ika Syntax*. Arlington TX: Summer Institute of Linguistics and the University of Texas at Arlington.
Givón, T. 1981. Typology and functional domains. *Studies in Language* 5: 163–193.
Gordon, L. 1986. *Maricopa Morphology and Syntax*. Berkeley CA: University of California Press.
Guirardello, R. 1999. *A Reference Grammar of Trumai*. PhD Dissertation, Rice University, Houston, Texas.
Haan, J.W. 2001. *The Grammar of Adang: A Papuan Language Spoken on the Island of Alor, East Nusa Tenggara – Indonesia*. PhD Dissertation, University of Sidney. [http://www-personal.arts. usyd.edu.au/jansimps/haan/adang-index.htm (November 13, 2006)].
Hagman, R.S. 1977. *Nama Hottentot Grammar*. Bloomington IN: Indiana University Press.
Hardman, M.J. 2000. *Jaqaru*. München: Lincom Europa.
Harris, A. 1981. *Georgian Syntax: A Study in Relational Grammar*. Cambridge: CUP.
Haspelmath, M. 2006. Against markedness (and what to replace it with). *Journal of Linguistics* 42: 25–70.
Hawkins, J.A. 2004. *Efficiency and Complexity in Grammars.* Oxford: OUP.
Heath, J. 1980. *Basic Materials in Warndarang: Grammar, Texts and Dictionary*. Canberra: Australian National University.
Henderson, J. 1995. *Phonology and Grammar of Yele, Papua New Guinea*. Canberra: Australian National University.
Iwasaki, S. & Ingkaphirom, P. 2005. *A Reference Grammar of Thai*. Cambridge: CUP.
Kakumasu, J. 1986. Urubu-Kaapor. In *Handbook of Amazonian Languages*, Vol. 1, D.C. Derbyshire and G.K. Pullum (eds), 326–403. Berlin: Mouton de Gruyter.
Kimball, G.D. 1991. *Koasati Grammar*. Lincoln NB: University of Nebraska Press.
King, G. 1993. *Modern Welsh: A Comprehensive Grammar*. London: Routledge.
Kittilä, S. 2005. Optional marking of arguments. *Language Sciences* 27: 483–514.
Kouwenberg, S. 1994. *A Grammar of Berbice Dutch Creole*. Berlin: Mouton de Gruyter.
Kusters, W. 2003. *Linguistic Complexity: The Influence of Social Change on Verbal Inflection*. PhD Dissertation, University of Leiden.
Kutsch Lojenga, C. 1994. *Ngiti: A Central-Sudanic Language of Zaire*. Köln: Rüdiger Köppe Verlag.
LaPolla, R.J. with Huang, C. 2003. *A Grammar of Qiang*. Berlin: Mouton de Gruyter.
Lindström, E. 2002. *Topics in the Grammar of Kuot, a Non-Austronesian Language of New Ireland, Papua New Guinea*. PhD Dissertation, Stockholm University.
Maddieson, I. 2006. Correlating phonological complexity: Data and validation. *Linguistic Typology* 10: 106–123.
McGregor, W. 1990. *A Functional Grammar of Gooniyandi*. Amsterdam: John Benjamins.
McWhorter, J. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology* 5: 125–156.
Miestamo, M. 2005. *Standard Negation: The Negation of Declarative Verbal Main Clauses in a Typological Perspective*. Berlin: Mouton de Gruyter.
Miestamo, M. 2007. Symmetric and asymmetric encoding of functional domains, with remarks on typological markedness. In *New Challenges in Typology: Broadening the Horizons and Redefining the Foundations*, M. Miestamo & B. Wälchli (eds), 293–314. Berlin: Mouton de Gruyter.

Murane, E. 1974. *Daga Grammar, from Morpheme to Discourse*. Norman OK: SIL.
Næss, Å. 2004. What markedness marks: The markedness problem with direct objects. *Lingua* 114: 1186–1212.
Nagaraja, K.S. 1999. *Korku Language: Grammar, Texts, Vocabulary*. Tokyo: Tokyo University of Foreign Studies.
Nakayama, T. 2001. *Nuuchahnulth (Nootka) Morphosyntax*. Berkeley CA: University of California Press.
Nichols, J. 1992. *Linguistic Diversity in Space and Time*. Chicago IL: The University of Chicago Press.
Payne, D.L. & Payne, T.E. 1990. Yagua. In *Handbook of Amazonian Languages*, Vol. 2, D.C. Derbyshire & G.K. Pullum (eds), 249–474. Berlin: Mouton de Gruyter.
Penchoen, T. 1973. *Tamazight of the Ayt Ndhir*. Los Angeles CA: Undena Publications.
Quintero, C. 2005. *Osage Grammar*. Lincoln NB: University of Nebraska Press.
R Development Core Team 2006. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. [http://www.R-project.org].
Rice, K. 1989. *A Grammar of Slave*. Berlin: Mouton de Gruyter.
Romero-Figueroa, A. 1997. *A Reference Grammar of Warao*. München: Lincom.
Rounds, C. 2001. *Hungarian: An Essential Grammar*. London: Routledge.
Saeed, J.I. 1999. *Somali*. Amsterdam: John Benjamins.
Schaub, W. 1985. *Babungo*. London: Croom Helm.
Shosted, R.K. 2006. Correlating complexity: A typological approach. *Linguistic Typology* 10: 1–40.
Sneddon, J.N. 1996. *Indonesian: A Comprehensive Grammar*. London: Routledge.
Sridhar, S.N. 1990. *Kannada*. London: Routledge.
Terrill, A. 2003. *A Grammar of Lavukaleve*. Berlin: Mouton de Gruyter.
Valenzuela, P.M. 1997. *Basic Verb Types and Argument Structures in Shipibo-Konibo*. MA Thesis, University of Oregon.
Vennemann, T. 1973. Explanation in syntax. In *Syntax and Semantics*, Vol. 2, J. Kimball (ed.), 1–50. New York NY: Academic Press.
Vries, L. de 1993. *Forms and Functions in Kombai, an Awyu Language of Irian Jaya*. Canberra: Australian National University.

Assessing linguistic complexity

Patrick Juola
Duquesne University

The question of "linguistic complexity" is interesting and fruitful. Unfortunately, the intuitive meaning of "complexity" is not amenable to formal analysis. This paper discusses some proposed definitions and shows how complexity can be assessed in various frameworks. The results show that, as expected, languages are all about equally "complex," but further that languages can and do differ reliably in their morphological and syntactic complexities along an intuitive continuum. I focus not only on the mathematical aspects of complexity, but on the psychological ones as well. Any claim about "complexity" is inherently about process, including an implicit description of the underlying cognitive machinery. By comparing different measures, one may better understand human language processing and similarly, understanding psycholinguistics may drive better measures.

1. Introduction

Most people with any background in language have at least an informal understanding of language complexity: a language is "complex" to the extent that you have to study in order to pass the exam on it, and in particular to the amount of stuff you simply have to memorize, such as lists of irregular verbs, case systems and declension patterns, and apparently arbitrary aspects of words such as gender. But to compare languages objectively requires a more formal specification of complexity, ideally one suited to a unified numerical measurement. Many ad hoc complexity measures have been proposed, of which Nichols (1986) is an obvious example; she counts the number of points in a typical sentence that are capable of receiving inflection.

McWhorter's (2001, 2005) definition encompasses a number of similar ad hoc measures (e.g., a language is more complex if it has more marked members in its phonemic inventory, or if it makes more extensive use of inflectional morphology), but he ties this, at least in theory, to a single numerical measure, the length of the grammar that a language requires. Despite the obvious practical difficulties (how do you compare two different inflectional paradigms, or how do you balance simple morphology with complex phonology), this provides a reasonable formulation for judging complexity.

However, as will be discussed in the remainder of this chapter, the question of "length" itself raises the issue of in what language (in a mathematical sense, i.e., set of primitive operations) the description should be made. It is argued here that information

theory provides several different approaches that yield different answers to questions of "linguistic complexity", and that analysis of data from natural languages can shed light on the psychological and cognitive processes appropriate to such description.

2. Information theory basics

2.1 Zipf and Shannon

The clearest statement of the motivation of the current work on an information-theoretic basis can be found in Zipf (1965 [1949]: 19–20), in his argument about applications of words:

> Man talks in order to get something. Hence man's speech may be likened to a set of tools that are engaged in achieving objectives. True, we do not yet know that whenever man talks, his speech is invariably directed to the achievement of objectives. Nevertheless, it is thus directed sufficiently often to justify our viewing speech as a likely example of a set of tools, which we shall assume to be the case. Human speech is traditionally viewed as a succession of words to which "meanings" (or "usages") are attached. We have no quarrel with this traditional view which, in fact, we adopt. Nevertheless in adopting this view of "words with meanings" we might profitably combine it with our previous view of speech as a set of tools and stated: words are tools that are used to convey meanings in order to achieve objectives. . . .
>
> Now if we concentrate our attention upon the possible internal economies of speech, we may hope to catch a glimpse of their inherent nature. Since it is usually felt that words are "combined with meanings" we may suspect that there is latent in speech both a more and a less economical way of "combining words with meanings," both from the viewpoint of the speaker and from that of the auditor.

Information theory provides a way, unavailable to Zipf, of resolving the two viewpoints he distinguishes. The speaker's economy requires that he express messages in the most compact form possible, up to a single word that can be used in any situation for any meaning. The hearer's economy requires that the speaker be easily understandable, and thus that the amount of message reconstruction effort (including the effort of listening to the statement) be minimized. A certain minimum amount of information must be conveyed in the speaker's signal so that a listener can distinguish his messages, but at the same time, too much useless information will clog things up. One can thus see that both the speaker and hearer have incentives to make the channel as efficient and easily understood as possible.

This framework can be placed on a firm mathematical footing (Shannon 1948; 1951). Shannon analysed all communications as a series of messages along a channel between an information source and a listener, and established mathematical bounds

for the maximum amount of information, measured in bits, that can be carried along such a channel. If less than this information is sent, some messages will not be distinguishable from each other. If more is sent, the "extra" is wasted resources. This is, for any source, a measure of the information content of that source and a lower bound on the amount of time/space/bandwidth necessary to send messages from that source and be understood. And in particular, to achieve this optimum use requires a Zipf-like framework, where the most frequently sent messages have the least "information" associated with them.

This can provide a framework for mathematical analysis of the intuitive notion of language complexity. All language is communication between the speaker and hearer; a message is "complex" if it has a large information content, and a language is "complex" if sending the message in that language requires much more bandwidth than the information content of the message.

In practical terms, there are a number of problems with the direct application of Shannon's formalism. It requires the researcher to enumerate beforehand all possible "messages" that might be sent along with their probabilities. But more serious problems lurk in the theoretical underpinnings. In the simplest formulation, Shannon's message probabilities are independent, meaning that the chance of a message being sent does not change, depending upon what other messages have been sent. In natural language, context matters, and understanding a message may depend crucially on the content of the previous messages. Finally, the assumption that the receiver may need to be able to recover the message perfectly might be problematic, since the knowledge I wish to transmit may be less than the language structure demands; if all I wish to say is that one of my siblings has red hair, the sex of the relevant sibling is irrelevant, as is their comparative age, but some languages (e.g., Japanese) may force me not only to specify the sex of my sibling, but whether they are older or younger, or to engage in a long, roundabout, marked, complex paraphrasing. In short, the "messages" of Japanese are different than those of English, suggesting another approach may be more fruitful.

2.2 Kolmogorov and other complexity definitions

Another important measurement of complexity is that of Kolmogorov complexity (Li & Vitányi 1997). Kolmogorov complexity measures the informativeness of a given string (not, as in Shannon's formulation, a message source) as the length of the algorithm required to describe/generate that string. Under this formulation, a string of a thousand alternating 'a's and 'b's would be easily (and quickly) described, while a (specific) random collection of a thousand 'a's and 'b's would be very difficult to describe. For a large corpus of messages, this could be used as an operationalisation of the average amount of information contained per message. In particular, notice that for any given set of messages produced by a Shannon source, it is a very efficient use of the channel to transmit, instead of a stream of individually coded messages, an algorithmic generator for the specific stream of interest. This illustrates the close relationship

between Shannon entropy and Kolmogorov complexity: Shannon's entropy is an upper bound on (and asymptotically equal to) Kolmogorov complexity.

Although the mathematics required to prove this is non-trivial, the result can be seen intuitively by observing that a decompression program and a compressed file can be used to (re)generate the original string.

Unfortunately, Kolmogorov complexity is formally uncomputable, in a strict technical sense related to the Halting Problem. Despite this technical limitation, Kolmogorov complexity is of interest as an unattainable ideal. If, as argued above, Kolmogorov complexity represents the ultimate possible file compression, a good file compressor can be seen as an attempt to approximate this kind of complexity within a tractable formal framework. By restricting the kind of operations permitted, it may be possible to develop a useful complexity measurement framework.

One example of such a restriction is that of linear complexity (Massey 1969; Schneier 1996). Linear complexity addresses the issue by assuming that the reconstruction machinery/algorithm is of a specific form, a linear feedback shift register composed of an ordered set of (shift) registers and a (linear) feedback function. (LFSR, Figure 1).

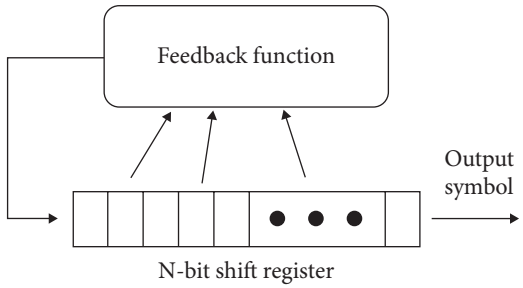


Figure 1. Sample Linear-Feedback Shift Register (LFSR).

The register set acts as a queue, where the past few elements of the sequence line politely up in order, while the feedback function generates the next single text element and adds it to the end of the queue (dropping the element at the head, of course). The linear complexity of a given sequence is defined as the size of the smallest LFSR generating a given sequence, and can be efficiently determined by a simple computer program (Massey 1969). Such systems are widely used as cryptographic encoding systems and random number generators because of their tremendous ability to generate long, apparently complex sequences from very little information.

What are the implications of this form of complexity? In this framework, the next element of the sequence (be it a word, morpheme, phoneme, etc.) is completely determined by a deterministic function of the most recent N elements of the sequence. As will be discussed later, there is no place in this framework for the notion of context, lexicon, or even long-term memory.

Another commonly-used framework is that of Ziv-Lempel complexity (Lempel & Ziv 1976; Ziv & Lempel 1977), the complexity metric that underlies most of the ZIP

family of commercial file compressors, which by contrast involves long-term memory almost exclusively. In this framework, the system builds a collection (lexicon?) of previously-seen strings and compresses new strings by describing them in terms of previously seen items. This adaptive "lexicon" is not confined solely to traditional lexemes, but can also incorporate short-range collocations such as phrasal verbs or common combinations. It has been proven that Ziv-Lempel complexity, as instantiated in LZ77 (Ziv & Lempel 1977) or similar techniques will, given infinite computing power and memory, compress any communications stream to any desired closeness to the theoretical maximum density, without prior knowledge of the messages or probabilities of the information sources. The popularity of the ZIP program is a testament to the practical effectiveness of this model.

At least in theory, this argument (although not necessarily the psychological underpinnings) would apply to any other proposed form of file compression.

These techniques provide a method of testing and measuring the amount of information and the amount of redundancy in any string. In particular, as will be discussed in the following section, by comparing the analytic results of a complexity metric on a large set of comparable messages, one can infer detailed analysis not only of overall complexity, but of its linguistic and cognitive components.

3. Linguistic Experiments

3.1 Information and language

To perform this analysis, we first must consider the types of information of interest. Elsewhere (Juola 1997), it has been argued that the information relevant to a text can be broken down into four major categories: the complexity of the idea(s) conveyed, the complexity of the author's style, the complexity mandated by the language in which the author writes, and the shared information omitted between the author and her audience

The third aspect is of course what is traditionally meant by "linguistic complexity"; by holding other aspects constant and varying the language in which a particular communication occurs, this can be observed numerically as an overall language complexity metric. Neglecting for a moment the question of individual authorial style, we can compare this by measuring the information contained in several expressions of the same ideas. A language with high "complexity" in McWhorter's sense will require more measured information in samples written in that language.

As an example, consider a simple propositional message, such as 'my brother has red hair'. From a purely semantic analysis of the meanings of the words, it is apparent that a single person is being spoken of and that the person is neither myself, nor the listener. It is not necessary to transmit this explicitly, for example via third person singular verb inflection. The "information" in this inflection is therefore additional complexity

demanded by the structure of the language in which it is expressed. We can further distinguish between redundancy and complexity in the nature of this information. If English (or any language) were perfectly regular, the nature of the inflection would be perfectly (and redundantly) predictable from the semantics expressed. In a more realistic setting, the inflection carries additional information necessary to describe the inflection itself (e.g., the irregularity of the verb to have) that is irrelevant, unnecessary, and "complex." In theory, a perfect compression program would be able to extract all the redundant information, but would still need to track the complexity.

Other examples of this sort of mandatory complexity would include gender markings, various forms of agreement, different lexical paradigms such as conjugations or declensions, and different syntactic structures such as word order, passivisation, and so forth. We can continue this analysis by focussing on the information contained in specific aspects or levels of language. For example, a language that is morphologically rich would be expected to contain much of its mandatory information in the system of inflectional morphology. By systematically distorting the samples to remove inflectional morphology, one can determine the importance of this type of complexity in the overall measure.

3.2 Validating complexity measurements

A first requirement for any empirical test of these theories is a suitable testbed; a collection of the same messages (as far as possible) expressed in different languages. Fortunately, this material is widely available in translated literature; the Bible, in particular, is an ideal sample (Resnik, Olsen & Diab 1999). It is relatively large, widely available, for the most part free of copyright restrictions, and generally well-translated.

In conjunction with the definitions of the previous section, we can then formalize common wisdom (pace McWhorter) regarding linguistic complexity as follows: For any complexity metric, the measured complexity of the Bible should be the same regardless of the language or translation.

A preliminary study (Juola 1998) has provided some data in support of this hypothesis. Using Bible versions in six languages (Dutch, English, Finnish, French, Maori, and Russian), it was shown that the variation in size of the uncompressed text (4242292 bytes, +/– 376471.4, or about 8.86% variation) was substantially more than the variation in size after compression via LZ (1300637 bytes, +/– 36068.2, about 2.77%). This strongly suggests that much of the variance in document size (of the Bible) is from the character encoding system, and that the underlying message complexity is (more) uniform.

We can contrast this with the results found (Juola 2000) by a comparable study of the linear complexity of the Bible translations. In this study, it was shown that the linear complexity of the Bible varies directly with the number of characters in the relevant translation, and that there is no descriptive advantage to the linear complexity framework. (We are tempted to draw the related conclusion that this similarly illustrates that the underlying process behind linear complexity does not describe human cognition well.)

Ideally, we would have some way of directly validating compression-based complexity measurements. One approach to this has been by setting up artificial environments such as inflectional paradigms and to vary the complexity of the paradigms. One such experiment (Juola, Bailey & Pothos 1998) developed an artificial lexicon and systematically varied the percentage of words subject to a simple inflectional "rule," from 0 to 100%. The resulting word list(s) were subjected to an operation the researchers termed "scrampression," repeated permutation followed by compression using the LZ formulation. The results, replicated here as Figure 2, show that, as expected, the "simplest" system is no inflection at all, that the system where all words were inflected was more complex (reflecting the additional complexity of the rule itself), but that the measured complexity varied smoothly and continuously, and that the most complex system (as predicted by Shannon's mathematics) was at intermediate stage where the question of whether a word was subject to the rule was itself complex.

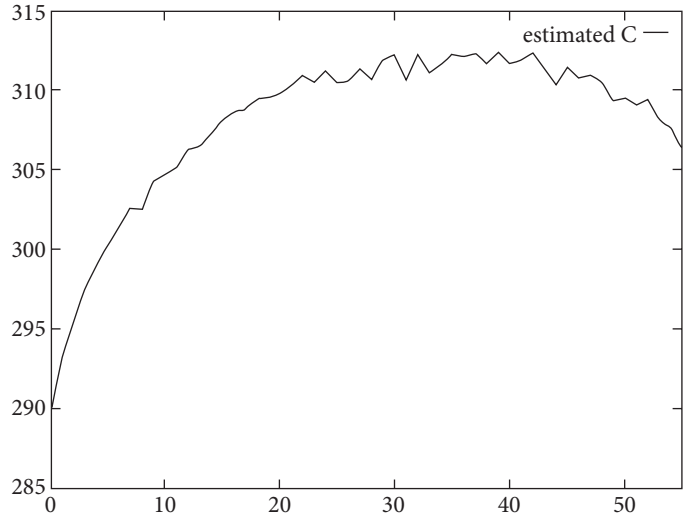


Figure 2. Variations in measured complexity as number of inflected words varies from 0 (0%) to 55 (100%).

From a mathematical and methodological perspective, these results indicate that LZ-based complexity measurements are practical to take from estimated appropriate corpora, and give meaningful measurements. Linear complexity measurements can also be taken, but do not map to our preliminary intuitions and cannot be validated against the world. From a psycholinguistic perspective, this argues strongly that LZ, with its focus on the lexicon and long-term storage and retrieval of words is a better model of the underlying regularities of language as expressed in corpora than linear complexity is. This is of course unsurprising; no sane person would believe a theory of language processing that did not include a lexicon. But this also suggests that careful choice of a model (and comparison of results) can illustrate other aspects of the

cognition underlying human language, either in ways that it is the same or different from traditional, information-theory based computer programs. In particular, there are lots of other non-LZ based compression programs, many of which apparently work better on language than LZ-compression. Can we infer properties of the human mind from these programs?

3.3 Further compression analyses: Translation

The previous section discussed three positive factors in linguistic complexity in conjunction with an apparent negative aspect, "the complexity/information omitted as shared between the author and her audience." One observation made by scholars of the translation process is that text in translation tends to be more explicit than the source material. This is easy to understand from a theoretical perspective. Information believed to be held in common between the speaker and hearer need not be expressed. For example, a writer and a reader who share a common knowledge of the city layout can be less explicit in giving directions than a writer and her out-of-town visitor.

One of the more obvious examples of common knowledge is, of course, the (shared) knowledge of a language as held between two speakers. Attached to this formal knowledge, in addition, is a host of other associations, cultural cues, and general knowledge that can be broadly lumped together as "cultural context." It is reasonable to assume that two fluent speakers of the same language/dialect share a tremendous amount of associations and general knowledge that is not necessarily a formal aspect of the language; just as a physicist can omit the definition of "momentum" when talking to another physicist, a native Englishman can omit the observation that Manchester is north of London when talking to another Englishman, but not necessarily when speaking to a foreigner. Similarly, these cities' locations would probably be known to a British reader but not necessarily to the readers of the same work in translation. Baker (1993) suggests that this feature of increased specificity may be a universal aspect of the translation process and presents a brilliantly clear example, where the (English) sentence "The example of Truman was always in my mind" is translated, into Arabic, into a huge paragraph – the English translation of this paragraph containing four sentences and ninety-one words, explaining Truman's relationship to the American Presidency and the end of the Second World War, his political background, and the metaphorical relevance of this "example." This sort of implicit information need not even necessarily be "factual," but can instead be habits of thought, expectations, associations, and so forth.

If true, this conjecture and theoretical framework may allow us to measure the degree of shared knowledge between a reader and writer, even between the writer and a "general audience." For the simplest case, that of two different linguistic communities viewed as separate "general audiences," the primary difference is, of course, the social and cultural context of both. This conjecture has been tested directly (Juola 1997) using the LZ compression test defined above.

In particular, the Biblical data defined above incorporate the original language as one of the test set.[1] If the translation process adds information as suggested above, the Hebrew original should be significantly shorter than any of the translations after compression.

Table 1. Sizes and information content (in bytes) of Pentateuch.

Language	Size (raw)	Size (compressed)
English	859937	231585
Dutch	874075	245296
Finnish	807179	243550
French	824584	235067
Maori	878227	221101
Russian	690909	226453
Mean	822485	235775
Deviation	70317	10222
HEBREW	(506945)	(172956)

As can be seen in Table 1, this is true.

Again, we see that the variation in compressed size is much smaller than the original, and we see that Hebrew is substantially smaller than either. Of course, written Hebrew omits vowels, but we can see that this is not the primary cause of this by observing other texts.

Table 2 shows similar sizes for a variety of translations of Orwell's *1984*, originally written in English and made available in a variety of languages by the Copernicus 106 MULTEXT-East Project or by the ECI/MCI corpus from the ACL.[2] Of the nine versions of 1984, the two smallest are the (original) English. Similarly, the smaller version of Dark Freedom (again distributed by ECI/MCI) is the Uzbek original.

Table 2. Information content of other works.

Work	Language	Size
1984	ENGLISH (v2)	228046
1984	ENGLISH	232513
1984	Slovene	242088
1984	Croat (v2)	242946
1984	Estonian	247355
1984	Czech	274310

(*Continued*)

1. The Bible as a whole is actually somewhat problematic in this regard as there is no single original language. For this reason, we momentarily restrict our attention to the first five books, which are all originally written in Hebrew.

2. This project is described at http://nl.ijs.si/ME/Corpus/mte-D21M/mte-D21M.html

Table 2. Continued.

Work	Language	Size
1984	Hungarian	274730
1984	Romanian	283061
1984	Bulgarian	369482
Dark Freedom	UZBEK	80209
Dark Freedom	English	90266

We thus see that the original version of corpora tend to be smaller than their translations, in keeping with the theoretical predictions of the translation scholars.

3.4 Complexity via distortion: Preliminary experiments

With this framework in hand, we are able to address more specific questions, such as the comparative role of syntax and morphology in the complexity of a particular language. In particular, just as one could measure the complexity of a morphological rule by systematically varying the degree to which it applied, so one can distort the morphology of a language sample as a whole to see the effects of "morphological complexity."

As before, we start with some definitions. In particular, we treat the ability to predict parts of a word token as a measure of morphology. For example, a plural context predicts a plural noun, which in English is typically marked with the *-s* suffix. The fact that the suffix *-ing* often signals a present participle and therefore something likely to follow the string "I am" is a morphological regularity. If the noun form is suppletively irregular, then there is no morphological complexity involved, simply a different lexical item. A morphologically complex language, under this view, is simply one where the information conveyed by these processes contributes substantively to the information conveyed by the entire text: for example, one where the agent/patient relationships cannot be determined by examination of the word order but can be determined by inspection of (e.g.) the regularities in the endings of the individual words.

By systematically distorting these regularities, we can measure the role of morphological information within a corpus. For example, rearranging the word order would destroy syntactic relationships (one would no longer be able to identify agent and patient in English), but not morphological ones (they would still be identifiable in strongly case-marked languages). Morphological regularities can be easily hidden by simple type-substitution.

jump	walk	touch	8634	139	5543
jumped	walked	touched	15	4597	1641
jumping	walking	touching	3978	102	6

Figure 3. Example of morphological degradation process.

Consider Figure 3: by replacing each type in the input corpus with a randomly chosen symbol (here a decimal number), the simple regularity between the rows and columns

of the left half of the table have been replaced by arbitrary relationships; where the left half can be easily described by a single row and column, the entire right half would need to be memorized individually. More accurately, this process is carried out token-wise where each token of a given type is replaced by the same (but randomly chosen) symbol which is unique to each type. This represents a hypothetical alteration to a language where the morphological rules have been "degraded" into a situation where all versions of related words can only be treated as suppletive forms, or alternately where the morphological regularity has been drastically altered in favor of a very complex system of lexical choice.

Note, however, that this does not eliminate lexical information. If, in the original text, "I am" predicted a present participle, the new symbols that replace "I am" will predict (the set of) symbols which correspond to and replace present participles. However, because these symbols have been randomly rewritten, there will be no easy and simple properties to determine which symbols these are. Examining Figure 3 again, the third row contains the present participles. In the left hand side, all entries in this row are instantly recognizable by their *-ing* morphology; the right hand side has no similar test. In other words, there is no longer a regular (morphological) way of predicting anything about the new token. Compressing the resulting substituted file will show the effects on the "complexity," i.e., the role of morphological complexity.

Performing this substitution has shown (Juola 1998) that languages differ strongly in the way and to the degree that this substitution affects their (compressed) sizes. The results are attached as Table 3.

Table 3. Size (in bytes) of various samples

Language	Uncompressed	Comp. (raw)	Comp. ("cooked")	R/C Ratio
Dutch	4509963	1383938	1391046	0.994
English	4347401	1303032	1341049	0.972
Finnish	4196930	1370821	1218222	1.12
French	4279259	1348129	1322518	1.01
Maori	4607440	1240406	1385446	0.895
Russian	3542756	1285503	1229459	1.04

As can be seen, the resulting r/c ratios sort the languages into the order (of increasing complexity) Maori, English, Dutch, French, Russian, Finnish. It is also evident that there is significant (phonological) information which is destroyed in the morphological degradation process, as three of the six samples actually have their information content *reduced*.

Validation of these numbers, of course, might be problematic; although few would argue with any measure of morphological complexity that puts Russian and Finnish near the top (and English near the bottom), there are few numbers against which to compare this ranking.

Fortunately, three of the languages in this sample (English, French, and Russian) are also part of the sample addressed in Nichols (1992). Nichols' finding is that English is less complex than either Russian or French, which are themselves equivalent.

The ranking here agrees with Nichols' ranking, placing English below the other two and French and Russian adjacent. This agreement, together with the intuitive plausibility, provides a weak validation.

Further examination of the corpora indicates some other interesting findings. These findings, presented as Table 4, can be summarized as the observation that (within the studied samples) the ordering produced by the complexity-theoretic analysis is identical with the ordering produced by the number of word types in the samples, and identical-reversed with the ordering produced by the number of word tokens. (This identity is highly significant; Spearman's rank test yields $p < 0.0025$, while Kendall's T test yields $p < 0.001$.) In other words, languages which are morphologically complex tend to have a wide variety of distinct linguistic forms, while languages which are morphologically simple have more word tokens, repeating a smaller number of types, as in Table 5.

Table 4. R/C ratios with linguistic form counts.

Language	R/C	Types in sample	Tokens in sample
Maori	0.895	19301	1009865
English	0.972	31244	824364
Dutch	0.994	42347	805102
French	1.01	48609	758251
Russian	1.04	76707	600068
Finnish	1.12	86566	577413

Table 5. Type/Token ratio across cased/uncased languages.

Concept/Case	English	Latin
Nominative	(the) night	nox
Genitive	of (the) night	noctis
Dative	to (the) night	nocti
Accusative	into (the) night	noctem
Ablative	in (the) night	nocte
Number of types	5(6)	5
Number of tokens	9(14)	5

This finding is intuitively plausible; the linguistic constructs which in language like Russian or Latin are expressed in terms of morphological variation are expressed in other languages through function words, which almost by definition are few types but many tokens. However, this finding is not necessary; it is theoretically possible for a language to exist that makes use of an extraordinarily wide variety of function word types

(thus inflating the number of types) or that inflates the number of tokens (for example by indicating plurality with repeated tokens). This finding then is further evidence for the approximate equality of overall linguistic complexity, at least within this sample.

3.5 Complexity via distortion: Subsequent experiments

My analysis of twenty-four Bibles, including nine English translations and sixteen non-English versions, shows, again, that languages are about equally complex, but that they express their complexity differently at different levels. (Table 6 shows the languages represented in this study.)

Table 6. Bible samples and languages used.

English Versions	Non-English Versions
American Standard Version (asv)	Bahasa Indonesian (bis)
Authorized (King James) Version (av)	Portuguese (brp)
Bible in Basic English (bbe)	Haitian Creole (crl)
Darby Translation (dby)	French (dby)
Complete Jewish Bible (jps)	Finnish (fin)
Revised Standard (rsv)	Hungarian (karoli)
Webster's Revised (rweb)	Dutch (lei)
Young's Literal Translation (ylt)	German (lut)
	Modern Greek (mgreek)
	Modern Greek (mgreeku) (unaccented)
	French (neg)
	Russian (rst)
	German (sch)
	German (uelb)
	Ukrainian (ukraine)

Methodologically, this study was slightly different, both the scope of the levels studied, but also in how distortion was performed. All samples were divided identically into verses (e.g., Genesis 1:2 represented the same "message" as best the translators could manage). Samples were distorted morphologically by the deletion of 10% of the letters in each verse at random. Samples were distorted syntactically by the deletion of 10% of the words (maximal non-blank sequences), while pragmatic distortion was obtained by deleting 10% of the verses themselves at random. We justify this as an exploration of "pragmatic" complexity by noting that sentences themselves must be interpreted through context. For example, the use of pronouns hinges on the availability of antecedents; languages where subjects can be dropped altogether make even heavier use of context. Just as elimination of words can distort the structure of individual sentences and make them difficult to follow, so eliminating previous sentences or verses, while having no effect on the syntactic acceptability of the individual sentences, will make the pragmatics of the discourse difficult to follow.

Although all three procedures will provably delete the same number of expected characters from the source files, the effects of compression on these files was expected to differ as argued above. Compression was performed using the UNIX gzip utility.[3]

Table 7 shows the compressed and uncompressed sizes (in bytes) of the various Biblical translations; Table 8 presents the (compressed) sizes both of the original and of the various distorted versions. Table 9 normalises all results by presenting them as multiples of the original size (e.g., the morphologically distorted version of 'asv' is 1.17 times its original size when both are compressed). This table is also sorted in order of decreasing value, and English versions are labelled in ALL CAPS for ease of identification.

Table 7. Uncompressed and compressed sizes (bytes) of various Biblical translations

Version	Uncompressed	Compressed
ASV	4280478	1269845
AV	4277786	1272510
BBE	4309197	1234147
bis	4588199	1346047
brp	3963452	1273549
crl	4351280	1312252
DBY	4227529	1265405
drb	4336255	1337768
fin	4169287	1370042
JPS	4309185	1288416
karoli	3957833	1404639
lei	4134479	1356996
lsg	4252648	1347637
lut	4180138	1341866
mgreek	4348255	1468263
mgreeku	4321200	1341237
neg	4188814	1323919
rst	3506401	1269684
RSV	4061749	1253476
RWEB	4247431	1262061
sch	4317881	1417428
uelb	4407756	1383337
ukraine	3564937	1315103
YLT	4265621	1265242

3. Another analysis, using a slightly different compression scheme (bzip2, using the Burrows-Wheeler transform), obtained similar results but will not be discussed further.

Table 8. Biblical sizes (bytes).

Ver.	Normal	Morph.	Prag.	Syn.
ASV	1269845.00	1487311.21	1162625.76	1235685.14
AV	1272510.00	1483580.35	1164269.82	1236878.18
BBE	1234147.00	1461058.00	1130314.66	1207880.69
bis	1346047.00	1601727.60	1229751.14	1282227.44
brp	1273549.00	1452517.41	1164547.00	1226345.93
crl	1312252.00	1518413.44	1200910.23	1285958.47
DBY	1265405.00	1474721.85	1158317.63	1229267.18
drb	1337768.00	1557772.69	1224481.67	1295321.43
fin	1370042.00	1548445.81	1251576.37	1294915.38
JPS	1288416.00	1504061.39	1179304.44	1253465.97
karoli	1404639.00	1575233.97	1283223.98	1330816.55
lei	1356996.00	1522644.85	1239991.16	1297103.54
lsg	1347637.00	1560146.19	1233249.90	1300432.48
lut	1341866.00	1529869.88	1226982.62	1285579.01
mgreek	1468263.00	1701221.45	1343161.70	1408603.16
mgreeku	1341237.00	1550600.30	1226936.07	1292009.44
neg	1323919.00	1533468.52	1211325.57	1277609.72
rst	1269684.00	1410919.44	1161101.89	1204901.46
RSV	1253476.00	1445795.94	1147694.76	1215960.88
RWEB	1262061.00	1471718.69	1154952.77	1227624.14
sch	1417428.00	1604828.85	1295267.52	1353695.77
uelb	1383337.00	1598355.09	1265317.06	1329331.17
ukraine	1315103.00	1451550.59	1201829.66	1247197.05
YLT	1265242.00	1482038.11	1158189.17	1230766.62

Table 9. Normalized sizes (bytes) after various transformations.

ukraine	1.10375	0.913867	0.948365
rst	1.11124	0.914481	0.948977
karoli	1.12145	0.913561	0.947444
lei	1.12207	0.913777	0.955864
fin	1.13022	0.913531	0.945165
sch	1.13221	0.913815	0.955037
lut	1.14011	0.914385	0.958053
bep	1.14053	0.914411	0.962936
RSV	1.15343	0.91561	0.970071
uelb	1.15543	0.914685	0.96096
mgreeku	1.1561	0.914779	0.963297
crl	1.15711	0.915152	0.979963
lsg	1.15769	0.91512	0.964972
neg	1.15828	0.914954	0.965021

(*Continued*)

Table 9. Continued.

mgreek	1.15866	0.914796	0.959367
drb	1.16446	0.915317	0.968271
DBY	1.16541	0.915373	0.971442
AV	1.16587	0.91494	0.971999
RWEB	1.16612	0.915132	0.972714
JPS	1.16737	0.915313	0.972874
ASV	1.17125	0.915565	0.973099
YLT	1.17135	0.915389	0.972752
BBE	1.18386	0.915867	0.978717
bis	1.18995	0.913602	0.952587

Examining these tables together supports the hypotheses presented earlier. First, all languages appear to be more uniform in size in their compressed than in their uncompressed sizes. Second, languages appear to differ in a reliable way in the degree to which their complexity is expressed in morphology or syntax; languages with high measured morphological complexity (as measured via morphological distortion) have low measured syntactic complexity (as measured via syntactic distortion). A new finding of this experiment is that all languages appear to be about equal in their pragmatic complexity; the variation in column 3 of Table 9 is almost nonexistent. This suggests that conversation-level constructions are handled similarly in all languages studied, perhaps reflecting underlying similarities in overall cognitive processing.

One minor point needs to be addressed. It is intuitively plausible that deleting words or phrases from a corpus should result in the overall lowering of information. It is less plausible that deleting letters should result in the overall increase in information, but that is a clear result of Table 9. The explanation is simple, but unhelpful. Essentially, by deleting letters randomly, the computer is creating a number of variant spellings of the same word, or alternatively a number of freely varying synonyms for each lexical item. Unsurprisingly, the compression program needs to determine (and store) which of the variant spellings is used in each instance. Strongly inflecting languages such as Finnish and Hungarian, ironically, suffer the least penalty because they have more word forms, and therefore fewer new variants are introduced in each one. Thus, the results show both that Finnish and Hungarian have high morphological complexity but low syntactic complexity, a fact that is both intuitive and supported by prior experiments.

It should finally be noted that in this sample, there is a single example of a creole language present. Sample 'crl' is Haitian Creole. Under McWhorter's hypothesis, this should be the smallest in compressed size. It is, in fact, about average. Readers who wish to regard this as evidence against McWhorter's hypothesis are sharply reminded of the dangers of drawing inferences from a single data point. As discussed in the next section, a much larger experiment on this scale should be run.

4. Discussion

The results presented above can be viewed as an extended proof for the idea that "language complexity" is a meaningful concept subject to empirical testing. In particular, the terminology and concepts of information theory can provide a structure for developing such studies. Taking a functionalist view of language – language serves to pass a message between the speaker and hearer – one can compare the mathematical properties of several different expressions of the same message.

In particular, by focussing on different translations of "the same" text, we can establish whether there are underlying differences in these properties that characterise the language in which the texts are written. By systematically distorting a single base text, we can establish whether this distortion introduces systematic changes in the properties.

The key questions are thus: what are the (mathematical) properties that can be measured, and how do they relate to the (linguistic) properties of general interest? In particular, how do our intuitive understandings of "complex" relate both to language and to mathematics?

The variety of definitions of complexity, both linguistic and mathematical, suggest that this relationship is not well-understood within either community. One possible approach to resolving this is to examine more closely the notion of process and the relationship of the processes captured in both psycholinguistics and the mathematical formalisms. In particular, most complexity formalisms start from a set of mathematical or algorithmic "primitives," and describe the complexity of an object in terms of the number of primitives required for it to function. For example, the fundamental primitives underlying Kolmogorov complexity are exactly those that underlie Turing machines: reading and writing unanalyzed symbols in an unbounded memory space. The fundamental primitives of LZ complexity include storage and retrieval from a learned "lexicon" of expressions. The "complexity" is defined as the number of primitives in a description of the process, ignoring completely the time element (how many primitives need actually be performed). By contrast, linear complexity assumes a very strong bound on the amount of memory available, and measures the complexity by the size of this very bound. Since human memory is known to have a large capacity, it makes intuitive sense that linear complexity should not well measure aspects of human cognition, while the plausibility of the lexicon strengthens the plausibility of LZ-based complexity measurements.

The framework developed in this paper thus provides researchers with a new way of investigating theories of the psycholinguistic processing of language. Specifically, by identifying the theoretical "primitives" implicit in the psychological theory, and developing a text compression method based upon those primitives, the results of compressing should be at least as plausible and at least as successful as the results obtained by other compression methods such as LZ-compression.

Similarly, a good theory of language processing should be able to capture the underlying regularities of human language in a sufficiently practical way that will be useful for the quite practical problem of text compression. With the ever increasing amount of text available (Nerbonne 2005), the ability to store ideas in a way that matches the efficiency of the human mind is almost certainly a desirable goal.

Finally, examination of algorithmic representation in terms of linguistic primitives may provide medical and psychological advantages as well. If our theory represents human performance under ideal conditions, studying how performance changes when theory is tweaked slightly may illustrate or explain aspects of human performance under less than ideal conditions, including both situational and cognitive degradation. This may improve science's ability both to diagnose and to treat linguistic/cognitive disorders.

5. Future work and conclusions

Briefly to recap the findings of the previous sections: the question of "linguistic complexity" can be formalized objectively and with relative precision using information theory. Depending upon the exact definition of "complexity" used, different measurements will be obtained. In particular, by using multiple translations of the same basic text (for example, the Bible or the novel 1984), one can see whether "all languages are about equally complex" as received wisdom requires, whether a particular language or language group is systematically less complex as McWhorter suggests, and whether there is any interesting variation or pattern in complexity.

Performing this experiment shows that languages do appear to be about equally complex when compared under the Ziv-Lempel definition of complexity. Less psycholinguistically plausible definitions, such as linear complexity, show correspondingly less plausible results. In particular, the variation in text lengths of uncompressed versions of the Bible was significantly greater than the variations in compressed lengths.

Further studies showed that varying particular aspects of language complexity (the regularity or irregularity of morphology) could similarly be detected by compression-based information theory measurements. Well-known aspects of language-in-translation, specifically the observation that translated texts need to be more explicit so that the reader can understand them against a different cultural background, can be observed by noting that translated texts are more complex, in this framework, than their originals. Finally, by systematically distorting particular types of expression, one can measure the role that type of expression plays in the overall complexity of language. Using this framework, studies have shown that languages differ in their morphological and/or syntactic complexity, and in particular, that languages with high morphological complexity have low syntactic complexity and vice versa. This is compatible with the tool-and-effort framework of Zipf and of Shannon; a language capable of sending information via morphology need not encode the same information syntactically.

A further finding – that languages do not appear to differ in their pragmatic complexity – may reflect the underlying cognitive universals of the human mind/brain system.

These findings only scratch the surface of what could be explored under this framework. In particular, McWhorter's theory that creoles are less complex than older languages could be tested by collecting a corpus of texts written in a variety of languages including both creoles and non-creoles. More carefully, one would want a text sample, written in one language and independently translated into a second (creole) and third (non-creole) language. Assuming that the same concepts are expressed in each translation, any significant size differences could only be attributed to the structural properties of the languages. If the creole translations were systematically smaller than the non-creole ones, this would be strong evidence supporting McWhorter's theoretical observations.

Another area of immediate research could be in the study of other formalisations of complexity and their usefulness and plausibility as linguistic measurements. The three concepts presented above are only a few of the dozens of definitions proposed and used by information theorists as compression technology. A systematic study of how different measurements vary would help researchers begin to understand which methods give intuitively plausible results and which do not.

Beyond this, the primary difference between the intuitively plausible results of the Ziv-Lempel metric and the implausible ones of linear complexity have been argued to be a result of the difference in handling of long-term memory and the lexicon. Can we find other methods that illustrate or demonstrate other important aspects of cognition?

This is the ultimate promise of complexity studies. A recent problem in linguistics has been dealing with the results of corpus-oriented studies; the problem of inferring linguistic processes from simply collections of products. The process-oriented analysis implicit in information theory may provide a useful tool to help in this inference and bridge the gap between the paper and the mind.

References

Baker, M. 1993. Corpus linguistics and translation studies: Implications and applications. In *Text and Technology: In Honour of John Sinclair*, M. Baker, G. Francis & E. Tognini-Bonnelli (eds), 233–250. Amsterdam: John Benjamins.

Juola, P. 1997. A numerical analysis of cultural context in translation. In *Proceedings of the Second European Conference on Cognitive Science*, Manchester, UK, 207–210.

Juola, P. 1998. Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics* 5 (3): 206–213.

Juola, P. 2000. A linear model of complexity (and its flaws). In *Fourth International Conference on Quantitative Linguistics* (QUALICO-2000), Prague, Czech Republic, R.H. Baayen (ed.), 7–8.

Juola, P., T.M. Bailey & E.M. Pothos 1998. Theory-neutral system regularity measurements. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society (CogSci-98)*, Madison WI, M.A. Gernsbacher and S.J. Derry (eds), 555–560.

Lempel, A. & J. Ziv 1976. On the complexity of finite sequences. *IEEE Transactions on Information Theory* IT-22 (1): 75–81.

Li, M. and P. Vitányi 1997. *An Introduction to Kolmogorov Complexity and Its Applications*. 2nd edn. Graduate Texts in Computer Science. New York: Springer.

Massey, J.L. 1969. Shift-register synthesis and BCH decoding. *IEEE Transactions in Information Theory* IT-15(1): 122–127.

McWhorter, J.H. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology* 6: 125–166.

McWhorter, J.H. 2005. *Defining Creole*. Oxford: Oxford University Press.

Nerbonne, J. 2005. The data deluge. *Linguistic and Literary Computing* 20(1): 25–40.

Nichols, J. 1986. Head-marking and dependent-marking grammar. *Language* 62: 56–117.

Nichols, J. 1992. Linguistic Diversity in Space and Time. Chicago IL: University of Chicago Press.

Resnik, P., M.B. Olsen & M. Diab 1999. The Bible as a parallel corpus: Annotating the 'Book of 2000 tongues'. *Computers and the Humanities* 33 (1–2): 129–153.

Schneier, B. 1996. *Applied Cryptography: Protocols, Algorithms and Source Code in C*. 2nd ed. New York: John Wiley and Sons, Inc.

Shannon, C.E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27 (4): 379–423.

Shannon, C.E. 1951. Prediction and entropy of printed English. *Bell System Technical Journal* 30 (1): 50–64.

Zipf, G.K. 1965 [1949]. *Human Behavior and the Principle of Least Effort*. New York NY: Hafner Publishing Company.

Ziv, J. & A. Lempel 1977. A universal algorithm for sequential data compression. IEEE Transactions on Information Theory IT23 (3): 373–343.

How complex are isolating languages?

David Gil
Max Planck Institute for Evolutionary Anthropology

How complex are isolating languages? The *Compensation Hypothesis* suggests that isolating languages make up for simpler morphology with greater complexity in other domains, such as syntax and semantics. This paper provides detailed argumentation against the Compensation Hypothesis. A cross-linguistic experiment measuring the complexity of compositional semantics shows that isolating languages rely more heavily on simple *Associational Semantics*, in which the interpretation of a combined expression is maximally vague or underdifferentiated, anything having to do with the interpretations of the constituent parts. In addition, it is argued that such vagueness is not necessarily resolved via recourse to context and a more complex pragmatics. Thus, it is concluded that isolating languages may indeed be of greater overall simplicity that their non-isolating counterparts.

1. Introduction

How complex are isolating languages, those characterized by minimal morphology, with little or no affixation or other kinds of word-internal structure? Do such languages compensate for the paucity of morphology by greater elaboration in other domains of grammar, or are they instead really simpler than other, non-isolating languages? This paper argues that in some cases, at least, isolating languages may indeed be of less overall complexity than their non-isolating counterparts.

The received view amongst linguists is that all languages are of equal complexity: there is no such thing as a simple language or a more complex one. However, in recent years, this view has been called into question, as more and more scholars have begun to entertain the possibility that languages may in fact differ from one another with respect to their overall levels of complexity: witness, for example, the various chapters in the present volume.

One obvious challenge faced by the received equal-complexity view is that within individual subsystems, languages clearly and undeniably vary enormously with respect to complexity. Rotokas has only 6 consonants in its phonemic inventory while !Xóõ has 122 (Maddieson 2005: 10); Vietnamese has exactly one inflected form for each of its verbs while Archi has 1,502,839 (Kibrik 1998: 467); English has but a single conventionalized speech level as defined in terms of specific vocabulary and morphological

items while Javanese has at least 5 (Errington 1998). Thus, in order to uphold the equal-complexity position, it is necessary to assume that languages compensate for lesser complexity in one area with greater complexity in another (or, alternatively, make up for greater complexity in one domain with lesser complexity in another). This assumption is referred to here as the *Compensation Hypothesis*. Various perspectives on the Compensation Hypothesis, including some critical ones, may be found in recent work by Shosted (2006), Fenk-Oczlon and Fenk (this volume), and Sinnemäki (this volume).

In the context of isolating languages, the Compensation Hypothesis suggests that languages with simpler morphology make up for their putative deficiency with greater complexity in some other linguistic domain. One specific application of the Compensation Hypothesis to languages of the isolating type is that of Riddle (this volume), who argues that in Southeast Asian languages such as Hmong, Mandarin and Thai, morphological simplicity is counterbalanced by greater elaboration in the lexicon and other grammatical domains. In principle, in accordance with the Compensation Hypothesis, morphological simplicity could be made up for by complexity in any other linguistic domain, from phonetics through to pragmatics. Most commonly, however, it is assumed that isolating languages compensate for their simpler morphology by means of increased complexity in one particular grammatical domain, namely, syntax.

A well-known example of this assumption can be found in the traditional view of the historical development from Latin to the modern Romance languages. Consider the following two examples of a Latin sentence in (1/2a), and its French equivalent in (1/2b):

(1) a. *Octavianus Athenis venit*
Octavian.NOM Athens.ABL come.PFV.3SG
b. *Octavien est venu d' Athènes*
Octavian be.PRS.3SG come.PST.PTCP from Athens
'Octavian came from Athens.'

(2) a. *Octavianus Cleopatram amabat*
Octavian.NOM Cleopatra.ACC love.IPFV.3SG
b. *Octavien aimait Cléopâtre*
Octavian love.IPFV.3SG Cleopatra
'Octavian loved Cleopatra.'

A key feature of this development is the drift from the more complex morphological structures of Latin towards the simpler morphological structures characteristic of the modern Romance languages. Thus, in the above examples, the Latin noun-phrases are marked for morphological case, while their French counterparts are not. Crucially, however, there are syntactic differences as well. In (1), in the oblique noun-phrase 'Athens', the Latin ablative suffix is replaced by the French preposition *de* (reduced to *d'*). Analogously, in (2), the distinction between nominative and accusative case suffixes in Latin gives way to a syntactic distinction in French: whereas in Latin the order of the words is flexible and they can be scrambled without change in basic meaning, in

French the order is fixed – if *Octavien* and *Cléopâtre* change places in the sentence, the meaning is reversed accordingly. What this shows, then, is that in the history of the Romance languages, the burden of differentiating thematic roles has shifted from the morphology to the syntax.

In accordance with the traditional view, these differences between Latin and modern Romance languages are not fortuitous; rather, they are a straightforward consequence of the Compensation Hypothesis (even if this hypothesis is not formulated explicitly in as many words). Thus, for example, with regard to facts such as in (1), Lehmann (1985: 312) writes that loss of case marking and the grammaticalization of prepositions "are in mutual harmony and favour each other." Similarly, with reference to data such as in (2), Kroch (2001: 701) asserts that "the loss of morphological case distinctions due to phonological weakening at the ends of words is generally thought to lead to rigidity of word order to compensate for the increase in ambiguity induced by the loss of case." Thus, according to this view, morphological and syntactic complexity balance each other out in systematic fashion: as the morphological complexity of case marking is lost, syntactic complexity is acquired in the form of prepositions and rigid linear order.

Although still widely held, the above view rests on shaky empirical foundations. For example, within the domain of case marking and word order, Kiparsky (1996) observes that Icelandic has case marking but rigid word order, while Enfield (to appear) shows that Lao has no case marking but flexible word order. Recent advances in linguistic typology have made it possible to test the Compensation Hypothesis against a much wider range of languages, and the results are a very mixed bag. For example, with respect to case marking and word order, Siewierska (1998: 509–512), making use of a sample of 171 languages, shows that there is a statistically significant correlation between case marking and flexible word order. However, she still finds plenty of languages which go against the correlation: 9 with case marking but rigid word order, and 5 with no case marking but the highest degree of flexibility of word order (allowing for 4 or more of the 6 possible orders of subject, object and verb).[1]

Thus, for morphology and syntax at least, there may indeed exist some kind of compensatory mechanism in accordance with which languages strive to attain a certain degree of expressive power, in order, for example, to distinguish between thematic roles, and that if one domain (e.g., morphology) fails to do the job, the other domain (e.g., syntax) steps in to pick up the slack. However, such a compensatory mechanism, if indeed it does exist, is relatively weak, even by the low standards characteristic of many of the implicational universals proposed in the field of linguistic typology. As evident from Siewierska's study, there are simply too many counterexamples in

1. Siewierska's study is concerned only with case marking on subjects and objects, not on obliques. Moreover, her definition of case marking includes clitics and adpositions, and thus would not distinguish between, say, the Latin ablative suffix in (1a) and the corresponding French preposition in (1b).

both directions – "fussy languages" with redundant marking of semantic distinctions in both the morphology and the syntax, as well as "carefree languages" that do not mark these distinctions in either domain – for such a compensatory mechanism to be attributed a significant role in determining linguistic structure.

This paper focuses on the latter class of counterexamples to the Compensation Hypothesis, those languages in which simplicity with respect to morphology is matched by an apparent simplicity in the domain of syntax. However, rather than examining the syntax of such languages in direct fashion, the spotlight is shifted from the syntax to a related semantic property. More specifically, this paper is concerned with the domain of *compositional semantics*, the ways in which the meanings of composite expressions are built up from the meanings of their constituent parts.

As a rule, linguistic forms bear meanings, and more complex forms bear more complex meanings. Accordingly, formal complexity, where form subsumes both morphology and syntax, correlates positively with complexity of compositional semantics. Compare, for example, Latin example (1a) with its ablative case-marked noun-phrase *Athenis* to its French counterpart (1b) with its prepositional phrase *d'Athène*. Morphologically the former is more complex while syntactically the latter is of greater complexity; however, in terms of overall formal complexity the two are equivalent. And indeed, the two are of similar complexity also with respect to their compositional semantics. In both examples, a lexical morpheme denoting a place occurs in construction with a grammatical morpheme denoting the abstract relationship of source to produce a composite expression whose meaning combines these two notions in the same straightforward way. Thus, for cases such as this, in which morphological complexity correlates negatively with syntactic complexity, the level of complexity of the compositional semantics should be similar in isolating and non-isolating languages. However, as suggested above, the negative correlation between morphological and syntactic complexity is far from universal.

This paper presents an empirical cross-linguistic exploration of the compositional semantics of isolating languages, comparing it with that of their non-isolating counterparts. The results of the paper show clearly that – contrary to the Latin/French case – the level of complexity of the compositional semantics varies across language types, and that in general, isolating languages tend to have simpler compositional semantics than non-isolating languages. The results of this paper thus cast further doubt on the Compensation Hypothesis as a viable principle governing the structure of languages. In particular, they suggest that at least some isolating languages may indeed be simpler than their non-isolating counterparts not just with respect to their morphology but also with regard to other structural domains.

2. Associational semantics

The complexity of compositional semantics in a given language may be measured with reference to a basic and very simple rule that lies at the heart of the compositional semantics of all languages: *associational semantics*.

The nature of associational semantics can be most readily understood with reference to an example drawn not from natural languages but rather from an artificial one, the language of pictograms, those iconic signs that can be found at airports, railway stations, the sides of roads, and other similar locations. Consider the following example:

(3)

Example (3) is formed by the juxtaposition of two simple signs, each with its own "lexical semantics". The arrow means something like 'thataway', 'over there', or 'go in that direction', while the sign to its right is a clear iconic representation of the meaning 'bicycle'. So much is straightforward; but what we are interested in here is the meaning of the example as a whole, that is to say, the interpretation of the juxtaposition of the two signs, its compositional semantics.

Readers in many European cities will be familiar with (3) as a semi-conventionalized traffic sign designating a special lane for bicycles. However, in other contexts, (3) can also be seen providing directions to a bicycle shop or exhibition. Thus, the above sign has a broad range of possible interpretations, including, among others, 'bicycles go thataway' and 'go thataway for bicycles'. Nevertheless, in (3), there is little or no evidence to support the existence of ambiguity involving two or more distinct interpretations. Rather, the construction is more appropriately characterized as vague with respect to its various possible interpretations, possessing, instead, a single general underspecified meaning, along the lines of 'something to do with thataway and bicycles'.

The notion of 'something to do with' may be put on a more rigorous footing with the following definition of the *Association Operator A*:

(4) *The Association Operator A*:
Given a set of n meanings $M^1 \ldots M^n$, the Association Operator A derives a meaning A ($M^1 \ldots M^n$), or 'entity associated with M^1 and . . . and M^n'.

The Association Operator is defined and discussed in Gil (2005a, b). Two subtypes of the Association Operator may be distinguished, the *Monadic Association Operator*, in which n equals 1, and the *Polyadic Association Operator*, for n greater than 1.

In its monadic variant, the Association Operator is familiar from a wide variety of constructions in probably all languages. Without overt morphosyntactic expression, it is manifest in cases of metonymy such as the often cited *The chicken left without paying*, where the unfortunate waiter uses the expression *the chicken* to denote the person who ordered the chicken. In other cases, the Monadic Association Operator is overtly expressed via a specific form, which is commonly referred to as a *genitive*, *possessive* or *associative* marker.

In its polyadic variant, the Association Operator provides a basic mechanism of compositional semantics in which the meaning of a complex expression is derived from the meanings of its constituent parts. In accordance with the Polyadic Association Operator, whenever two or more expressions group together to form a larger expression, the meaning of the combined expression is associated with, or has to do with,

the meanings of each of the individual expressions. Thus, when applying to example (3), the Association Operator assigns the juxtaposition of signs the interpretation A(THATAWAY, BICYCLE), or 'entity associated with thataway and bicycle': a unitary underspecified interpretation which, in the right context, can be understood as 'bicycles go thataway', 'go thataway for bicycles', or in any other appropriate way. In general, this is how the compositional semantics of pictograms works: when two or more signs are juxtaposed, the Polyadic Association Operator applies to produce a single underspecified interpretation associated in an indeterminate fashion with the interpretations of each of the individual signs. The language of pictograms may thus be characterized as possessing *associational semantics*.

So much for pictograms; but what of natural languages? As argued in Gil (2005a), there is good reason to believe that the Polyadic Association Operator is a universal cognitive mechanism underlying the compositional semantics of all languages. However, whereas with pictograms the Polyadic Association Operator stands alone as the sole means of deriving compositional semantics, in the case of natural languages it provides for a basic and very general interpretation which, in most cases in most languages, is subsequently narrowed down by a variety of more specific semantic rules making reference to any number of morphosyntactic features such as linear order, nominal case marking, verbal agreement, and so forth. Hence, in order to observe the Polyadic Association Operator in action, it is necessary to dig a little harder. Three possible domains in which one can search for associational semantics are *phylogenetically*, back in evolutionary time; *ontogenetically*, back in developmental time; and *typologically*, across the diversity of contemporary adult human languages; see Gil (2005a) for further details.

In the latter case, that of contemporary adult human languages, the Polyadic Association Operator constitutes the foundation for additional more specific rules of compositional semantics, making reference to a wide variety of morphosyntactic features such as linear order, case marking, agreement, and so on. Although many of these rules might themselves be universal, numerous others are specific to individual languages. Accordingly, languages may differ in the extent to which their compositional semantics makes use of such additional rules. This variation results in a typology of associational semantics. At one end of the scale are *compositionally associational languages*, in which the number of additional semantic rules is relatively few and their combined effect relatively insignificant, while at the other end of the scale are *compositionally articulated languages*, in which the number of additional semantic rules is relatively large and their combined import much more substantial.

As argued in Gil (2005a, b), Riau Indonesian provides an example of a compositionally associational language. Consider the following simple sentence in Riau Indonesian:

(5) *Ayam* *makan*
chicken eat

Sentence (5) might be glossed into English as 'The chicken is eating'; however, such a translation, or for that matter any other translation that one might come up with,

cannot but fail to adequately represent the wide range of interpretations that is available for the Riau Indonesian sentence, wider than that of any simple sentence in English. To begin with, the first word, *ayam* 'chicken', is unmarked for number and definiteness, while the second word, *makan* 'eat', is unspecified for tense and aspect. But the semantic indeterminacy is more far reaching yet. The thematic role of *ayam* 'chicken' is also unmarked: in addition to agent, it could also be patient ('Someone is eating the chicken'); moreover, given an appropriate context, it could assume any other role, for example benefactive ('Someone is eating for the chicken'), comitative ('Someone is eating with the chicken'), and so forth. Moreover, the ontological type of the entire construction is itself underspecified: in addition to an event, it could also denote an object ('The chicken that is eating'), a location ('Where the chicken is eating'), a time ('When the chicken is eating'), and so on. Thus, sentence (5) exhibits an extraordinarily large degree of semantic indeterminacy, in which many of the semantic categories encoded in the grammars of most other languages are left underspecified. Rather than being multiply ambiguous, sentence (5) is thus vague, with a single unitary interpretation, most appropriately represented in terms of the Polyadic Association Operator, as A (CHICKEN, EAT), or 'entity associated with chicken and eating'. As suggested by these facts, the compositional semantics of sentence (5) works just like the juxtaposition of signs in example (3). Accordingly, Riau Indonesian may be characterized as a compositionally associational language.

However, Riau Indonesian differs in an important respect from the language of pictograms: although largely compositionally associational, it still possesses a number of additional grammatical rules whose effect is to constrain the degree of semantic indeterminacy. To cite just one example, in (5) above, although *ayam* 'chicken' may assume any thematic role whatsoever, there is a significant preference for it to be understood as agent ('The chicken is eating') as compared to patient ('Someone is eating the chicken'). In Gil (2005b), facts such as these are accounted for in terms of additional semantic rules making reference to head-modifier structure, iconicity, and information flow. It is rules such as these which justify the characterization of Riau Indonesian as a largely rather than purely compositionally associational language.

Still, as evidenced by facts such as those in (5), the combined effect of these rules is much less far reaching than in other, more compositionally articulated languages. Thus, for example, in English, in a sentence such as *The chicken is eating*, there is obligatory grammatical encoding for all the semantic categories that are left unspecified in Riau Indonesian: number, definiteness, tense, aspect, thematic role, ontological type and others. In English, then, associational semantics is the foundation for a wealth of additional rules of semantic representation, most or all of which, quite naturally, make reference to the considerably greater morphosyntactic elaboration of English, in comparison to Riau Indonesian. Thus, with respect to their compositional semantics, languages may range from compositionally associational to compositionally articulated, in accordance with the extent to which the Polyadic Association Rule is supplemented with further, more specific rules of semantic interpretation.

3. Measuring complexity: The association experiment

The typology of associational semantics presented above provides a straightforward measure of the complexity of the compositional semantics of a given language. To wit, compositional semantics is simpler in compositionally associational languages, and more complex in compositionally articulated languages. Thus, the association typology provides a means of testing the Compensation Hypothesis and measuring the complexity of isolating languages. The Compensation Hypothesis predicts that isolating languages should resemble their non-isolating counterparts with respect to the association typology. Is this indeed the case? An answer to this question is provided by the *Association Experiment*.

The Association Experiment is designed to measure the degree to which languages are associational, that is to say, the extent to which their compositional semantics relies on the Polyadic Association Operator to the exclusion of other more specific rules of semantic interpretation.[2] The Association Experiment focuses on languages with the following two typological properties: (a) isolating; and (b) apparent SVO basic word order. The motivation for this choice of focus is as follows. To begin with, there is a priori reason to believe that, as a linguistic type, isolating languages exhibit greater diversity with respect to the associational typology than do other linguistic types. In agglutinating, synthetic and polysynthetic languages, the elaborate morphology typically supports many additional rules of semantic interpretation on top of the Polyadic Association Rule, resulting in a high degree of articulation of the compositional semantics. In contrast, isolating languages have fewer morphosyntactic devices at their disposal, and are therefore more likely to exhibit a range of values from compositionally articulated to compositionally associational.

Any attempt to measure, by means of objective experimentation, the extent to which a language is compositionally associational faces an immediate problem: how do we know which rules are responsible for a given construction being interpreted in a certain way if not through the mediation of linguistic analysis conducted by a linguist within a particular theoretical framework? The answer of course is that we don't, at least not for sure. However, at the level of prima facie plausibility, it seems possible to distinguish between interpretations probably attributable to the application of various specific rules of semantic interpretation making reference to morphosyntactic features such as linear order, case marking, agreement and the like, as opposed to interpretations which may more plausibly be attributed directly to the Polyadic Association Operator.

2. The construction of the Association Experiment would not have been possible without the assistance of Brad Taylor, who helped design the FileMaker Pro database, Yokebed Triwigati, who drew the beautiful pictures, and the many colleagues and friends who provided me with their expertise on individual languages, and helped me with obtaining subjects and with other technical matters.

Interpretations of the latter kind are referred to here as *Apparently Associational Interpretations*. The goal of the Association Experiment is thus to compare the availability across languages of such Apparently Associational Interpretations.

3.1 Experimental design

The Association Experiment is a truth conditional experiment, in which subjects are asked to judge whether certain states of affairs constitute possible interpretations of particular sentences. Each stimulus presents a written sentence in the target language, beneath which are two pictures. Subjects are asked to read the sentence and then look at the two pictures, after which they must respond to the question: Which of the two pictures is correctly described by the sentence? Subjects may provide any of four different responses: "the picture on the left", "the picture on the right", "both pictures", or "neither picture". In each stimulus, one of the two pictures is a designated *test picture*, whose availability as a possible interpretation for the given sentence is what is being tested by the experiment, while the other picture is an *alternative picture* whose availability as a possible interpretation is, for the most part, not of interest.

In total, the Association Experiment contains 44 stimuli; however, for the purposes of the present paper, we shall be concerned with just a subset of these, consisting of 16 stimuli. In these 16 stimuli, the designated test picture is contrasted with an alternative picture representing an interpretation that is completely impossible. A schematic representation of the 16 test stimuli is given in Table 1:

Table 1. Test sentences.

No	Stimuli	Construction type
1	CLOWN DRINK BOOK	
	Clown drinking while reading book	*Bare Peripheral Following*
2	CLOWN BUY HAPPY	
	Clown buying fruit with happy face	
3	CLOWN EAT RIVER	
	Clown eating by river	
4	SOLDIER CUT AXE	
	Soldier cutting wood with axe	
5	COFFEE LAUGH	
	Person spilling coffee, onlooker laughing	*Bare Peripheral Preceding*
6	MONEY HAPPY	
	Man holding money with happy face	
7	TABLE DANCE	
	People dancing on tables	
8	CHAIR JUMP	
	People jumping over chairs	

(*Continued*)

Table 1. Continued.

No	Stimuli	Construction type
9	BIRD EAT	
	Cat eating bird	*Bare Patient Preceding*
10	TIGER AFRAID	
	People fearing tigers	
11	DOG DRAW	
	Man drawing dog	
12	HAT SEW	
	Men sewing hats	
13	MOUSE CHASE CAT	
	Cat chasing mouse	*Bare Patient Preceding plus Agent*
14	MOUSE BITE SNAKE	
	Snake biting mouse	
15	CAR PUSH WOMAN	
	Woman pushing car	
16	ELEPHANT LOOK TIGER	
	Tiger looking at elephant	

In Table 1, each test stimulus is described in terms of the written sentence, represented schematically in small caps, and the test picture, for which a simple verbal description is given. (For reasons of space, the alternative picture is not represented.)

The schematic small-cap representation of the written sentence shows the expressions denoting basic concepts belonging to the major ontological categories of thing, property and activity, in the order in which they occur in the sentence. For the most part, it thus provides a representation of the content words common to all of the languages examined, to the exclusion of all of the additional grammatical items, lexical and morphological, which vary form language to language. For example, in stimulus 1, CLOWN DRINK BOOK represents, among others, the English test sentence *The clown is drinking the book*, and the Minangkabau test sentence *Badut minum buku*. Whereas for English, CLOWN DRINK BOOK leaves out the definite article *the* and the combination of auxiliary *is* and gerundive suffix *-ing* present in the test sentence, for a strongly isolating language such as Minangkabau, the same schematic representation happens to constitute a perfect interlinear gloss for the test sentence, consisting as it does of three content words with no additional grammatical marking. In a few cases, however, expressions denoting basic concepts may consist of more than a single content word. For example, in stimulus 1, CLOWN DRINK BOOK also represents the Bislama test sentence *Fani man i dring buk*, in which the basic concept CLOWN is expressed by means of a compound consisting of two content words *fani* 'funny' and *man* 'man'.

In each of the languages examined, the test sentences are the simplest natural-sounding and stylistically-neutral grammatical sentences allowed for by the language, in accordance with the schematic representations in Table 1. The languages examined

differ considerably with respect to the degree to which the shared content words must be supported by additional grammatical markers. Rather than being structural equivalents of one another, the test sentences in different languages might best be thought of as constituting "content-word equivalents" of each other, sentences making use of similar words arranged in superficially similar patterns. As to whether they are semantically equivalent: this is precisely what the Association Experiment is constructed to find out.

As indicated in the rightmost column of Table 1, the 16 stimuli divide into two groups of 8, each of which in turn divides into two groups of 4. The first group of 8 stimuli test for Apparently Associational Interpretations in constructions of the *Bare Peripheral* type. Bare Peripheral constructions involve the juxtaposition of an expression denoting an activity and another expression denoting an entity which cannot be construed as filling a slot in the semantic frame of the former, activity expression. The term "peripheral" is thus intended as an antonym of the term "core" as in "core argument", while the term "bare" takes cognizance of the fact that the peripheral expression bears no marking to indicate the semantic nature of its relationship to the activity. Consider, for example, stimulus 1, and, in particular, the collocation DRINK BOOK. As inanimate but solid objects, books have no place in the semantic frame of drinking; accordingly, BOOK is characterized as peripheral. In a compositionally articulated language with, say, rigid SVO order, subjects should judge the sequence CLOWN DRINK BOOK as semantically anomalous, since it requires an interpretation in which BOOK is the patient of DRINK; accordingly, they should reject the test picture of a clown drinking while reading a book. In contrast, in a compositionally associational language, subjects should be able to interpret the sequence CLOWN DRINK BOOK as A (CLOWN DRINK BOOK), that is to say, as meaning essentially anything involving a clown, a drinking, and a book, and in such a case, they should accept the test picture of a clown drinking while reading a book as a possible interpretation of the sentence.

Thus, stimuli 1-8 test for Apparently Associational Interpretations of the *Bare Peripheral* type. However, they do so in a number of distinct but related ways. Whereas in stimuli 1-4 the bare peripheral expression follows the activity expression, in stimuli 5-8 the bare peripheral expression precedes it. A further, somewhat more subtle distinction can be made with respect to the nature of the semantic relationship between the bare peripheral expression and the activity expression. In general, the semantic relationship is looser in stimuli 1, 2, 5, and 6 than it is in stimuli 3, 4, 7, and 8. Whereas in the former cases, the semantic relationship can only be characterized as associational, in the latter cases, the semantic relationship is of a type that is typically referred to as a subcategory of oblique, either locative or instrumental.

The second group of 8 stimuli test for Apparently Associational Interpretations in constructions of the *Bare Patient Preceding* type. Bare Patient Preceding constructions involve an expression denoting an activity preceded by an expression denoting an entity understood as the patient of that activity. Consider, for example, stimulus 9. In a

compositionally articulated language with rigid SVO order, subjects should interpret the sequence BIRD EAT as entailing that BIRD is the agent of EAT; accordingly, they should reject the test picture of a cat eating a bird, since the bird is assigned the wrong thematic role, that of patient. In contrast, in a compositionally associational language, subjects should be able to interpret the sequence BIRD EAT as A (BIRD EAT), that is to say, as meaning essentially anything involving a bird and an eating, hence, in particular, they should accept the test picture of a cat eating a bird as a possible interpretation of the sentence.

Thus, stimuli 9-16 test for Apparently Associational Interpretations of the *Bare Patient Preceding* type. In terms of S, V and O, these stimuli thus test for OV order (in languages that are supposed to be of SVO basic word order.) Stimuli 9-16 also fall into two distinct subgroups. Whereas stimuli 9-12 contain constructions involving only a patient expression followed by an activity expression, stimuli 13-16 contain constructions in which a patient expression and an activity expression are in turn followed by an expression referring to the agent of the activity.

3.2 Running the experiment

The Association Experiment is conducted on a laptop computer, making use of a database in FileMaker Pro. Subjects are tested individually, one after another. Subjects are presented with the stimuli one at a time, each stimulus occupying the entirety of the computer screen. Different subjects receive different randomized orders of the various stimuli. Each response, either verbal, such as "the one on the left", or gestural, such as pointing to the chosen picture, is recorded immediately by the experimenter, by clicking on a button located beneath the picture chosen, or, in the case of "both" or "neither" responses, between the two pictures. Whereas in some situations, I was able to conduct the experiment on my own, in other cases I was helped by a local assistant, whose role was to round up subjects and/or help me communicate with them.

The Association Experiment is an ongoing project, still in progress; by the end of 2006, over 1500 subjects had been tested in over 20 languages. For each language, the experiment is run on a population of subjects conforming to a *Baseline Sociolinguistic Profile*: uneducated, low-to-middle class, over 12 years of age, living in a community where the test language is spoken, and tested in their home community, in a public or semi-public area. In addition, within individual languages, the experiment is also run on additional populations outside the Baseline Sociolinguistic Profile.

This paper represents an initial report on a limited subset of the experimental results obtained so far. A total of ten languages are reported on: two non-isolating, English and Hebrew, and eight isolating. The eight isolating languages fall into four groups of two languages each: (1) creoles: Papiamentu and Bislama; (2) West African: Twi and Yoruba; (3) Southeast Asian: Cantonese and Vietnamese; and (4) West Indonesian: Minangkabau and Sundanese. For each language, data is presented from a population of subjects conforming to the Baseline Sociolinguistic Profile.

4. Results

The results of the experiment for the ten test languages considered in this paper, conducted on populations conforming to the Baseline Sociolinguistic Profile, are presented in Table 2.

Table 2. Results of the Association Experiment (Baseline Sociolinguistic Profile).

			Availability (%) of Apparently associational interpretations	
Language group	**Language**	**Number of subjects**	**Bare Peripheral**	**Bare Patient Preceding**
Non-Isolating	English	32	7	4
	Hebrew	30	10	4
Creole	Papiamentu	25	14	5
	Bislama	17	51	15
West African	Twi	21	22	12
	Yoruba	31	68	26
Southeast Asian	Cantonese	30	42	12
	Vietnamese	38	67	15
West Indonesian	Minangkabau	30	74	57
	Sundanese	35	76	49

In Table 2, the first two columns specify the language and the group to which it belongs. The third column shows the number of subjects conforming to the Baseline Sociolinguistic Profile who performed appropriately on the distractor items, and whose responses are therefore taken into account. The fourth and fifth columns present the results of the experiment: the availability, in percentages, of Apparently Associational Interpretations, averaging over the two major types of constructions, Bare Peripheral (stimuli 1-8) and Bare Patient Preceding (stimuli 9-16).[3]

The results in Table 2 show that for both of the non-isolating languages, English and Hebrew, the availability of Apparently Associational Interpretations is, as expected, very low, in the 0–10 % range. The results of the experiment accordingly support the characterization of English and Hebrew as compositionally articulated languages.

3. For the purposes of the above calculations, a response is taken to reflect the availability of an Apparently Associational Interpretation in either of two cases: if the subject chose the test picture, or if the subject chose both the test and the alternative picture. Due to subjects' general preference to choose exactly one of the two pictures, the first case is overwhelmingly more frequent than the second.

Thus, English and Hebrew provide a base point with respect to which the eight remaining languages, all of an isolating type, may be compared.[4]

Looking at the remaining eight isolating languages, one is immediately struck by the large amount of variation between them. With respect to the availability of Bare Peripherals, the figures range from 14% in Papiamentu to 76% in Sundanese, while with regard to the availability of Bare Patients Preceding, the figures range from 5% in Papiamentu to 57% in Minangkabau. Thus, in terms of the availability of Apparently Associational Interpretations, isolating languages are clearly not of a single type. Nevertheless, comparing the eight isolating languages to the two non-isolating languages, English and Hebrew, a striking pattern emerges. Except for Papiamentu, the availability of Apparently Associational Interpretations is consistently, and in some cases even dramatically higher in the isolating languages than it is in their two non-isolating counterparts.

Among the eight isolating languages, a systematic difference is evident in the relative availability of Apparently Associational Interpretations for the two construction types, with Bare Peripherals scoring significantly higher than Bare Patients Preceding in all languages. What this suggests is that in isolating languages, the domain of core arguments is more likely than other more peripheral domains to exhibit tighter grammatical structure, resulting in less flexible word order, and thereby setting the stage for construction-specific rules of compositional semantics making reference to linear order. In other words, core arguments tend to be more articulated, while peripheral expressions provide a more congenial environment for associational semantics. Nevertheless, even within the domain of core arguments, the availability of Apparently Associational Interpretations is generally higher in the eight isolating languages than in the two non-isolating ones.

The results of the association experiment thus cast serious doubt on the viability of the Compensation Hypothesis, at least within the domains under consideration here, namely, morphology, syntax and compositional semantics. As pointed out earlier, the Compensation Hypothesis predicts that isolating languages should make up for their simpler morphology with greater complexity in the syntax. Since compositional semantics can refer to either morphological or syntactic structures, the level of complexity of the compositional semantics should thus be roughly the same in isolating and non-isolating languages. In particular, in the case at hand, these two language types should resemble each other with respect to the association typology: the availability of Apparently Associational Interpretations should be the same for isolating and for non-isolating languages. However, the results of the experiment show that the availability of Apparently Associational Interpretations in isolating languages is systematically higher than it is in non-isolating languages. In other words, isolating

4. Since the results reported on in this paper represent work in progress, the figures have not yet been submitted to rigorous statistical analysis. Impressionistically, however, most of the generalizations discussed herein are so robust that it is very unlikely that they will not prove to be statistically significant.

languages tend to be more compositionally associational, while non-isolating languages tend to be more compositionally articulated. Thus, contrary to the predictions of the Compensation Hypothesis, isolating languages are actually characterized by simpler compositional semantics than their non-isolating counterparts.

The results of the association experiment accordingly shed new light on the relationship between morphology and syntax as exemplified in Section 1. Broadly construed, the two types of Apparently Associational Interpretations, Bare Peripheral and Bare Patient Preceding, correspond to the two pairs of Latin and French sentences in (1) and (2), respectively. The availability of Bare Peripheral interpretations in highly associational isolating languages shows that in such languages, oblique expressions such as those in (1) may occur with neither Latin-like case markings nor French-style adpositions. Similarly, the availability of Bare Patient Preceding interpretations in highly associational isolating languages shows that such languages may manage perfectly well with neither case marking as in Latin (2a) nor rigid word order as in French (2b).

Consider the following rough equivalents of (1) and (2) in the isolating and highly associational language Minangkabau:

(6) *Ujang tibo Pariaman*
Ujang arrive Pariaman
'Ujang arrived from Pariaman.'
(preferred: 'Ujang arrived at Pariaman.')

(7) *Kartini cinto Ujang*
Kartini love Ujang
'Ujang loved Kartini.'
(preferred: 'Kartini loved Ujang.')

In (6), the oblique expression *Pariaman* bears neither case nor adpositional marking. Accordingly, its semantic relationship to *tibo* is underspecified: the compositional semantics of the collocation *tibo Pariaman* is thus nothing more than that provided by the association operator, A (ARRIVE, PARIAMAN), or 'entity associated with arriving and Pariaman'. As indicated above, the preferred interpretation of *tibo Pariaman* is one in which *Pariaman* is the goal of *tibo*; however, it is also possible to understand it in such a way that *Pariaman* is the source of *tibo*. In (7), the two arguments, *Ujang* and *Kartini* bear no case or adpositional marking; moreover, the order in which they occur with respect to each other and to *cinto* can be changed freely without any effect on the range of possible interpretations. Here too, then, the thematic roles of the participants are underspecified, and the compositional semantics is associational: A (KARTINI, LOVE, UJANG), or 'entity associated with Kartini, loving and Ujang'. Again, as indicated above, the preferred interpretation of the sentence is that in which *Kartini* is the experiencer and *Ujang* the theme; however, it is also possible to understand it in such a way that *Kartini* is the theme and *Ujang* the experiencer. Thus, in both (6) and (7), the interpretations corresponding to those tested in the experiment (indicated directly beneath the interlinear gloss) are somewhat disfavoured relative to other available

interpretations, such as those corresponding to (1) and (2) (indicated in parentheses). It is due to such patterns of preferences that the availability of Apparently Associational Interpretations in Minangkabau, as evident in Table 2, remains well below the 100% mark. Nevertheless, as is equally evident in Table 2, the availability of Apparently Associational Interpretations in Minangkabau is very substantially above the near-0% mark characteristic of non-isolating languages: sentences (6) and (7) can indeed be understood in the same way as sentences (1) and (2) in Latin and French. Thus, what Minangkabau and other similar highly associational isolating languages show is that there is nothing at all inevitable about the apparent trade-off between morphological and syntactic complexity characteristic of the historical development from Latin to modern Romance languages such as French.

In summary, then, the results of the association experiment show that isolating languages tend to be at least as associational as their non-isolating counterparts, with some isolating languages exhibiting a substantially higher degree of compositional associationality. Thus, isolating languages do not compensate systematically for their minimal morphological structures with greater elaboration in the domain of syntax. Rather, isolating languages are not only morphologically simpler, they also tend to be simpler with regard to their compositional semantics. Accordingly, the results of the association experiment suggest that isolating languages may indeed be characterized by lesser overall complexity than other, non-isolating languages.

5. "But doesn't the pragmatics compensate?"

But here the plot thickens, and a new factor enters the fray: pragmatics. A widely held view suggests that languages with less developed morphosyntactic structures compensate with more elaborate rules of pragmatics. In accordance with this position, languages lacking overt formal expression of various semantic categories force their speakers to fill in the missing pieces, as it were, by appealing to linguistic and extralinguistic context. For example, in the domain of anaphora and the so-called "pro-drop parameter", Huang (1984) and others distinguish between "hot languages" such as English, in which pronouns are obligatory, and "cool languages" such as Chinese, in which they are optional, and in which, accordingly, speakers have to exercise their pragmatic competence in order to figure out the missing elements. Analogously, in the domain of thematic roles, Enfield (to appear) characterizes isolating languages such as Lao, and, in passing, also Mandarin, Thai, Vietnamese and Riau Indonesian, as "exemplifying the extreme of pragmatically-oriented grammar", asserting that "[w]hen core referential information is not symbolically encoded in grammar, potential ambiguities in role/reference relations are readily resolved by features of context."[5] Following such

5. In fact, in some of my earlier writing, I leaned towards a similar position with regard to the compositional semantics of Riau Indonesian, for example in Gil (1994: 195), where, after

reasoning, it is suggested that the burden is merely shifted from the morphosyntax to the pragmatics, while the overall level of language complexity remains the same. In particular, in the case at hand, compositionally associational languages would make up for their morphosyntactic simplicity with greater pragmatic complexity, and accordingly, their overall level of complexity would end up roughly the same as that of compositionally articulated languages.

One possible counterargument to this position would maintain that yes, indeed, impoverished morphosyntax is compensated by more elaborate pragmatics, but pragmatics does not really "belong" to the grammar, it is not part of the language, and therefore should not be taken into consideration when measuring overall complexity. To a certain degree, the issue is merely terminological: researchers are free to formulate hypotheses about overall complexity (or any other matter) with respect to grammar or language defined in any way whatsoever, provided of course that the relevant definitions are clear and unambiguous.

However, regardless of whether pragmatics is included within the scope of grammar or language, there is still a serious substantive problem with the position that isolating languages compensate for their simpler morphosyntax with more complex pragmatics. Specifically, such a position presupposes the existence of a universal meaning representational system with a fixed set of semantic features shared by each and every one of the world's languages, into which every utterance in every language must be decoded, if not by purely grammatical means then with the additional assistance of the pragmatics. But the existence of such a universal semantics is questionable on both practical and principled grounds.

In practical terms, when linguists say things like "the grammar doesn't encode it, so speakers have to figure it out from the context", the "it" in question is, more often than not, a distinction that is absent from the target language but one that the linguist expects to find, because it is prominent in English and other European languages. The mistake is to assume that this absence is a lacuna that needs somehow to be filled in by the pragmatics: in most cases, there is no such need whatsoever. As Becker (1995: 291) writes, "each language, from the point of view of another, appears full of holes." But the word "appear" is crucial here: the hole, supposedly to be filled in by the pragmatics, has no existence other than in the eye of the beholder.

Consider, for example, the category of number in Riau Indonesian. As pointed out above, in example (5) the word *ayam* 'chicken' is unmarked for number, and can be interpreted as either singular or plural. Does a speaker of Riau Indonesian, every time he or she hears the word *ayam*, wonder whether it refers to a single chicken or to a

---

presenting a precursor to the associational analysis outlined above with reference to example (5), I wrote that "the precise nature of the relationship between the constituent meanings, and who did what to whom, is determined by context and by extra-linguistic knowledge of the world." As will become evident below, I no longer adhere to this position.

plurality of chickens? In a few specific contexts probably yes, in a large majority of everyday situations surely no. To assert that the pragmatics of Riau Indonesian must compensate for the language's impoverished morphological number marking by forcing the speaker to figure out whether each and every word is meant to be singular or plural is tantamount to imposing an English grammatical category onto the semantics of Riau Indonesian. Analogous observations hold with respect to all of the other categories – definiteness, tense, aspect, thematic role, ontological type, and so forth – that are encoded in the grammars of English and other European languages but left unspecified in Riau Indonesian. Thus, when linguists say that the pragmatics of a language such as Riau Indonesian must work overtime in order to clear up the vaguenesses or ambiguities left behind by an excessively underdifferentiating grammar, they are lapsing into a prescriptive mode, asserting that all languages have to aim for the precise level of clarity and specificity that just happens to be characteristic of English and other familiar European languages.

It should be clear that the choice of English and other European languages as the basis for the putative set of universal semantic features is arbitrary and unmotivated. Consider, once again, the example of number. In Arabic, in addition to singular and plural there is also a dual number. Now if a Eurocentric linguist can suggest that in Riau Indonesian the pragmatics must disambiguate singular from plural, what is then to stop an Arabic grammarian from proposing that in impoverished English, where dual and plural are conflated, whenever a speaker encounters a plural noun, he or she must stop to figure out whether it is dual or three-or-more? But speakers of Arabic also have a problem, because in Larike (an Austronesian language of Indonesia), in addition to singular, plural and dual there is also a trial number. So would a Larike linguist have speakers of Arabic scrambling to determine whether each three-or-more noun is actually trial or four-or-more? (And in the meantime, our Riau Indonesian speaker is struggling to figure out whether *ayam* is singular, dual, trial, or four-or-more.) Once again, analogous observations hold with respect to just about any other grammatical category: if there is a language that is less grammatically differentiated than English with respect to the category in question, there is also bound to be one that is more grammatically differentiated than English with respect to it. Hence, a Eurocentric linguist cannot impose English-like semantic features on a language that does not encode such features without being willing to accept a similar imposition of more highly differentiated features from some other language onto English. And the result is a potentially infinite regress. What this shows, then, is that there is a serious practical problem with the presupposition of a universal meaning representational system with a fixed set of semantic features shared by each and every one of the world's languages: we simply do not know what these features are, and in our ignorance, opt for the convenient but unwarranted assumption that they are similar to those of English and other European languages.

However, in addition to the above practical problem, there is also a more principled reason to question the presupposition of a universal meaning representational

system with a fixed set of semantic features accessed by each and every one of the world's languages. To put it most simply, there is no a priori reason to assume that such a universal semantics should exist; and in the absence of such reason, the most appropriate default assumption is that no such universal semantics does exist. After all, languages differ from each other in just about every other domain, from phonetics through phonology, morphology and syntax to discourse structure. Also, the things that we talk about differ from language to language, be they cultural artefacts or aspects of the physical environment. In view of such variation, it would certainly be very surprising if it turned out that there did exist a meaning representational system and a set of semantic features that was invariant across languages. Of course, there will be lots that is universal in semantics, just as there are universal properties within every other domain of language, as well as commonalities shared by all cultures and physical environments; the point is that in addition to such universals we should expect to find lots of semantic features that one language encodes in its morphosyntax while another ignores, not only in its grammar but also in its rules of pragmatic inference. Thus, for example, in the absence of positive evidence to the effect that speakers of Riau Indonesian make regular and systematic use of their pragmatic competence to disambiguate singular, plural, dual, trial and so on, considerations of parsimony point to the conclusion that the meaning representational system for Riau Indonesian does not distinguish between categories of number.

In everyday parlance, vagueness has mildly negative connotations; however, vagueness is one of the central design features of linguistic semantics, and for good reason. The number of forms, grammatical or lexical, at our disposal, will always be several orders of magnitude fewer than the number of things we can imagine and then talk about with these forms. Thus, plural number will be vague with respect to the distinction between 234 and 235, a noun such as *chicken* will be vague with respect to dimensions such as weight, height, colour, texture of plumage, and so on and so forth. In some cases, we can get closer and closer to what we want to say by making use of the recursivity of syntax, narrowing down our message with more and more words. However, in other instances, we cannot even get close: this is what is called ineffability. Try describing the taste of a particular fruit to someone who has never experienced it. Human language is simply ill-equipped for such tasks. Vagueness is thus an inevitable consequence of the immense gulf between the huge number of things that we can conceptualize and the limited linguistic tools at our disposal to talk about them. For the most part, we do not notice this vagueness; it generally attracts our attention only in the comparative cross-linguistic context, when we encounter an unfamiliar language in which a distinction that we expect to find turns out not to be there. How can a language "do without" such and such a category, we wonder. But it is important to keep in mind that all languages are massively vague, and that in this context, the differences between languages such as Riau Indonesian and English are small indeed.

Less vagueness, or greater specificity, does not always entail greater functionality. Sometimes it is useful to have the right form at one's disposal, but in many other

cases, obligatory grammatical encoding constitutes an inconvenience. Vagueness is a necessary feature of language: it is there for a very good purpose, to expedite communication. Thus, it simply does not make sense to suppose that whenever a speaker encounters an instance of vagueness, he or she immediately makes use of pragmatic rules in order to fill in an arbitrary large number of additional semantic features. Of course, pragmatics is there to flesh out utterances with whatever additional information happens to be relevant in the given context. But it is not there in order to translate all utterances in all languages into a single universal meaning representation system with a shared set of semantic features.

In particular, it is not reasonable to assume that whenever a speaker of Riau Indonesian encounters a sentence such as *Ayam makan* in (5), he or she automatically embarks on some long and arduous pragmatic path of "figuring everything out": assigning number and definiteness to *ayam*, tense and aspect to *makan*, thematic role to *ayam*, and ontological type to the construction as a whole – essentially translating *Ayam makan* into English. Rather, the speaker simply fills in what is necessary in the situation at hand, and does not bother with anything else. Of course, in this respect, the speaker of Riau Indonesian is no different from other speakers dealing with other utterances in other languages.

In reaction to arguments such as these, it is sometime suggested that, well, alright, vagueness with regard to number, definiteness, tense and aspect is somehow liveable with, but surely thematic roles are an entirely different kettle of fish. While it is easy, so it is said, to imagine contexts in which it does not matter whether there is one chicken or several in the yard, surely it must always be of great concern whether our chicken is doing the eating or being eaten. How can a hearer of a sentence such as the Riau Indonesian *Ayam makan* not feel compelled to make whatever efforts are necessary to figure out the appropriate thematic role of the chicken? There is no denying that in most cases, speakers care whether chickens are eating or being eaten, or whether Octavianus loved or was loved. In such cases, if the language is isolating and highly associational, and the relevant thematic roles are not encoded in its grammar, speakers do indeed exercise their pragmatic competence and plumb the available linguistic and extra-linguistic context in order to fill in as much additional detail as is considered necessary. However, there is no reason to suppose that they go beyond what is necessary, in order to make distinctions that are not relevant to the given context, just because such distinctions are morphosyntactically encoded in some other language; see Gil (2001) for some specific examples of vagueness with respect to thematic roles in Riau Indonesian.

In conclusion, then, speakers of all languages make use of linguistic and extra-linguistic contexts in order to enrich semantic representations with further detail in a variety of domains, including number, definiteness, tense, aspect, thematic roles, and many others. However, such filling in of the details takes place only to the extent that it is required by the particular context at hand: its goal is not to translate utterances into some universal meaning representational system with a fixed set of semantic features.

Accordingly, there is little reason to believe that languages with less developed morphosyntactic structure compensate systematically for their grammatical simplicity with more complex rules of pragmatics.

6. Summary

How complex are isolating languages? This is the question that was posed at the beginning of this paper. In accordance with the Compensation Hypothesis, widely assumed by many linguists, isolating languages should make up for their simple morphology with complexity in some other domain. Most often, the domain in question is presumed to be syntax. However, there is already plenty of evidence showing that there does not exist any simple trade-off between morphological and syntactic complexity.

The results of the cross-linguistic association experiment presented in this paper demonstrate that isolating languages are actually simpler than their non-isolating counterparts with respect to their compositional semantics. The highly associational nature of isolating languages provides further evidence against the claim that they compensate for simpler morphology with more complex syntax. Moreover, it shows that the morphological simplicity of isolating languages is actually mirrored by simplicity in another, logically independent, linguistic domain, namely compositional semantics. Thus, there is prima facie reason to believe that isolating languages may indeed be of less overall complexity than other, non-isolating languages.

Nevertheless, the possibility exists that isolating languages may compensate for their simpler morphology by complexity in some other, completely different domain. One such proposal, suggesting that isolating languages shift the burden of expressive power in systematic fashion from the morphology to the pragmatics, was argued against in the preceding section, but in principle there could be any number of other potential compensatory mechanisms relating morphology to other linguistic domains. For example, Vietnamese, one of the isolating languages discussed in this paper, has an extremely rich inventory of vowels, tones and phonation types: perhaps, then, some or all isolating languages make up for their simple morphology with greater complexity in the domain of phonology. Although such scenarios cannot be ruled out, they are not very likely, and the burden of proof rests solidly on the shoulders of those who would wish to argue in support of such alternative compensatory mechanisms. In the meantime, in the absence of any such proposals, it seems reasonable to adhere to the conclusion that at least some isolating languages are indeed simpler than other, non-isolating languages.

Notwithstanding the titles of this paper and this volume, I remain agnostic as to whether the notion of complexity has an important role to play in the study of language, and, in particular, the field of linguistic typology. (In fact, I suspect that it may prove to be more relevant to domains such as phylogeny, diachrony, ontogeny and sociolinguistics than to the "straight" synchronic study of language.) The current state

of the art with regard to linguistic complexity is a rather odd one. As noted in the introduction, the conventional wisdom holds that all languages are of equal overall complexity; yet at the same time, most linguists seem to believe that the notion of overall complexity is somehow incoherent or ill-defined. However, these two views are clearly mutually contradictory: if you do not have an explicit and well-defined metric of complexity, you cannot go around saying that all languages must be of equal complexity – one of these two positions, at least, must be abandoned. This paper, like several others in the present volume, argues that the latter view, whereby all languages are of equal complexity, needs to be rejected. Still, this does not necessarily entail that complexity per se is an important, or "deep" design feature of language: it could just as easily be epiphenomenal. If, as argued in this paper, isolating languages are simpler than non-isolating languages, their simplicity need not necessarily be construed as a single holistic property setting them apart from other more complex languages. At least as plausible is a state of affairs in which the simplicity of isolating languages and the complexity of other non-isolating languages result from the accidental confluence of a wide range of logically independent properties, in morphology, syntax, semantics, and possibly other linguistic domains.

Abbreviations

ABL	ablative
ACC	accusative
IPFV	imperfective
NOM	nominative
PFV	perfective
PRS	present
PST	past
PTCP	participle
SG	singular
3	third person.

References

Becker, A.L. 1995. *Beyond Translation: Essays Towards a Modern Philology*. Ann Arbor: University of Michigan Press.

Enfield, N.J. (to appear). 'Case Relations' in Lao, a radically isolating language. In *The Handbook of Case*, A. Malchukov & A. Spencer (eds), Oxford: Oxford University Press.

Errington, J. 1998. *Shifting Languages: Interaction and Identity in Javanese Indonesia*. Cambridge: Cambridge University Press.

Gil, D. 1994. The structure of Riau Indonesian. *Nordic Journal of Linguistics* 17: 179–200.

Gil, D. 2001. Escaping eurocentrism: Fieldwork as a process of unlearning. In *Linguistic Fieldwork*, P. Newman & M. Ratliff (eds), 102–132. Cambridge: Cambridge University Press.

Gil, D. 2005a. Isolating-Monocategorial-Associational language. In *Categorization in Cognitive Science*, H. Cohen & C. Lefebvre (eds), 347–379. Oxford: Elsevier.

Gil, D. 2005b. Word order without syntactic categories: How Riau Indonesian does it. In *Verb First: On the Syntax of Verb-Initial Languages*, A. Carnie, H. Harley & S.A. Dooley (eds), 243–263. Amsterdam: John Benjamins.

Huang, C.-T.J. 1984. On the distribution and reference of empty pronouns. *Linguistic Inquiry* 15: 531–574.

Kibrik, A.E. 1998. Archi (Caucasian - Daghestanian). In *The Handbook of Morphology*, A. Spencer & A.M. Zwicky (eds), 455–476. Oxford and Malden: Blackwell.

Kiparsky, P. 1996. The shift to head-initial VP in Germanic. In *Studies in Comparative Germanic Syntax* 2, S. Epstein, H. Thráinsson & S. Peters (eds), 140–179. Dordrecht: Kluwer.

Kroch, A.S. 2001. Syntactic change. In *The Handbook of Contemporary Syntactic Theory*, M. Baltin & C. Collins (eds), 699–729. Malden: Blackwell.

Lehmann, C. 1985. Grammaticalization: Synchronic variation and diachronic change, *Lingua e Stile* 20: 303–318.

Maddieson, I. 2005. Consonant inventories. In *The World Atlas of Language Structures*, M. Haspelmath, M.S. Dryer, D. Gil & B. Comrie (eds), 10–13. Oxford: Oxford University Press.

Shosted, R. 2006. Correlating complexity: A typological approach. *Linguistic Typology* 10: 1–40.

Siewierska, A. 1998. Variation in major constituent order: A global and a European perspective. In *Constituent Order in the Languages of Europe, Empirical Approaches to Language Typology* [Eurotyp 20-1], A. Siewierska (ed.), 475–551. Berlin: Mouton de Gruyter.

Complexity in isolating languages: Lexical elaboration versus grammatical economy*

Elizabeth M. Riddle
Ball State University

Contrary to recent claims that highly analytic or isolating languages are simpler than synthetic languages, in large part due to lack of inflectional affixation in the former, I argue that although isolating Asian languages such as Hmong, Mandarin Chinese, and Thai may be economical in terms of inflection, they exhibit significantly more complex lexical patterns of particular types than more synthetic languages such as Polish and English in like contexts. Evidence includes the use of classifiers, reduplication, compounding, stylized four-part expressions, verb serialization, and other types of what I call "lexical elaboration." This analysis has implications for the question "What is linguistic complexity?" as well as for the more basic and vexing question of "What is grammar?"

1. Introduction

The widely accepted tenet among contemporary linguists that all human languages are equally complex is based on the observation of three cross-linguistic characteristics:

1. similarity in the rate of first language acquisition;
2. the potential to express propositions on any topic and of great conceptual complexity;
3. systematic, rule-governed structure involving hierarchical dependencies in addition to linear ordering.

Nonetheless, the doctrine of universal equivalent complexity is being re-examined, with calls for the development of criteria to measure complexity across languages.

One focus of discussion concerns the relative grammatical complexity of synthetic versus highly analytic or isolating languages. Dahl (2004) argues from an

* Special thanks are due to my Hmong teachers and consultants, Pheng Thao, Lee Thao, Lopao Vang and Leng Xiong, to Xiaojing Cheng and Apichai Rungruang, my Mandarin and Thai consultants, and to my husband Paul Neubauer for his tremendous assistance and support.

information-theoretic stance, separate from content expressiveness, that languages can differ in complexity of grammatical form. He proposes that complexity be measured according to the length of the description necessary to account for a given structure. In his view, "historically mature" (114) structures, such as inflection and obligatory syntactic rules, are more complex than structures simply involving juxtaposition and/or "verbosity" (2004: 52–53), i.e., the use of separate words to represent grammatical properties. Included among the relatively less complex structures are optional rules and marking. Thus, relatively analytic languages would be less grammatically complex than relatively synthetic languages, although no less complex in expressiveness. Parkvall (this volume) supports McWhorter's (1998, 2001a) claim that creoles, which notably lack inflection, are grammatically simpler than non-creoles. Responding to the observation by Mufwene (1986), among others, that Chinese, a non-creole, is also highly analytic but not considered by most linguists to be less complex than synthetic languages, McWhorter (1998, 2001a) argues that the complexities of tone and classifiers in Chinese distinguish it from creoles. The implication is that analytic languages like Chinese are more grammatically complex than creoles. McWhorter (2001b) asserts, however, in a response to DeGraff (2001), that creoles are not simple or primitive overall.

In contrast, based on experimental data from Riau Indonesian and other languages, Gil (2001, this volume) argues that even non-creole isolating languages are grammatically less complex than synthetic languages. He claims that the extent to which a language represents meaning associationally and allows multiple associations for a single form corresponds to the extent to which it is isolational. He concludes that such languages do not have a correspondingly more complex semantics to compensate for morphologically simpler structures.

With at least 500 relatively or highly analytic languages spoken by well over a billion of the world's people, such claims have major implications for the understanding of the nature of human language, linguistic change, cognition, and language-mediated social interaction.[1]

Using data from three Asian languages, (White) Hmong, Mandarin (Chinese), and (Bangkok) Thai (henceforth collectively referred to as HMT), I support the traditional view (Sapir 1921), that highly analytic languages are as grammatically complex as synthetic languages, although in different ways.[2] One aspect of this complexity is lexical

1. This is my own tentative calculation, based on data gleaned from the *Ethnologue* (Gordon 2005), supplemented by Hyman (2000), Matisoff (1986), Ramsey (1987), Ratliff (1992b), and Gil (this volume).

2. Hmong is usually considered to be Austro-Thai (Benedict 1975) or Sino-Tibetan. The Hmongic family is native to Laos, Thailand, Vietnam, Burma, and China. The White (Hmoob Dawb) and Green (Hmoob Ntsuab) varieties are also spoken by Laotian diaspora communities in the United States, Australia, and France, especially in (French) Guyana. This paper discusses

elaboration (Riddle 1992), i.e., the relative proliferation of free and compounded rather than bound morphemes as a form of lexical explicitness. Although HMT exhibit relative economy in the overt marking of grammatical relationships, they are typologically characterized by richness of syntactic and semantic representation via lexical means, both within the propositional semantics and in the lexicon. In contrast to Dahl's (2004) view of verbosity as a characteristic of simpler grammars, I argue that what might be called verbosity in HMT is a form of complexity. For example, grammatically free morphemes such as classifiers and aspect markers are extensively used to express syntactic and semantic relationships, analogous to the use of *to* as an infinitival marker in English *to buy*, in contrast to the morphological marking of Polish infinitives, as in *kupić* 'to buy' (perfective) and *kupować* 'to buy' (imperfective). Additionally, HMT lexical items are themselves lengthened and made more descriptively elaborate through compounding, and juxtaposition of closely related words is common.

2. Some typological characteristics of highly analytic languages

Foley and Olson (1985: 50–51) observe that the Kwa languages of West Africa share several interrelated structural properties with many languages of Southeast Asia, including phonemic tone, monosyllabicity, isolating morphology and verb serialization. They view these properties as having compensatory functional interrelationships, in that phonological loss can lead to the development of tone and the loss of affixes (generally unstressed), resulting in a tendency toward monosyllabicity, plus more rigid word order and verb serialization to counterbalance the loss of information provided by inflectional morphology. Riddle and Stahlke (1992) argue that the analytic Kwa languages and many mainland Southeast Asian (SEA) languages tend to disfavor syntactic explicitness in the sense of overt tagging or labeling, not only in terms of morphological inflection, but in other ways as well, as compared to relatively synthetic languages such as Modern Greek, Polish, or even English. Analytic languages generally have fewer different overt markers of syntactic dominance, and strongly favor compounding and reduplicating free morphemes in word formation as opposed to affixation, where minimally, one element is overtly morphologically dependent on another.

White Hmong (Hmoob Dawb), as spoken natively in Laos and Thailand. The data are in the Romanized Popular Alphabet, with hyphens inserted between syllables in compound words. Syllable-final consonants or lack thereof represent the tones of each syllable. The original orthography in (15) has been regularized. The tones are: b = high level (55); j = high falling (53); 0 final consonant = mid level (33); g = mid falling breathy (31); v = mid rising (34); s = low level (11); m = low checked (10); d = long low rising (213). (Tone numbers from Bisang 1996: 587). The Mandarin and Thai examples are given in standard transliterations with tone diacritics.

2.1 Complexity in HMT

In the following discussion, all major phenomena are prominent in HMT and many other languages of the region (Matisoff 1973). However, due to space limitations, not all phenomena will be illustrated for each language. Any data not otherwise attributed are from my own work with native speakers or my personal spoken competence.

2.1.1 *Classifiers*

McWhorter (1998) suggests that the classifier systems of many analytic East and Southeast Asian languages add to their grammatical complexity, distinguishing them typologically from creoles. I will show here how the HMT classifier systems are extremely complex in multiple ways, and suggest that the use of classifiers constitutes a form of lexical elaboration.

These classifier systems are typical for the region (Matisoff 1973; Conklin 1981; Bisang 1993). Most nouns are categorized for specific classifiers, similar to gender agreement in languages such as Greek and Polish. In contrast to a European type involving masculine, feminine, and possibly neuter classification, or even the Bantu categorization of nouns into as many as 15 classes (Lyovin 1997: 220–224), however, HMT make dozens of distinctions according to properties such as humanness, animacy, shape, amount, purpose, and individuation. Some classifications seem arbitrary, often rendered obscure by historical change, including metaphorical extension.

Heimbach (1969: 455–456) lists 76 classifiers for White Hmong. Li and Thompson (1981: 105) state that "Mandarin has several dozen classifiers." Conklin (1981: 43) describes Thai as having more than 100.

A small sample of Hmong classifiers is given in Table 1. The most central properties are listed, along with examples of nouns subcategorized for each. Similar types of categories also occur in Mandarin and Thai.

Table 1. Some classifiers in Hmong

CLF	Central Properties	Noun	Noun Meaning
leej	Humans (marked)	*neeg*	'person'
		txiv	'father'
tus	Animates (including humans);	*dev*	'dog'
	long, firm, inanimate objects	*cwj*	'stick'
txoj	Long, flexible objects;	*xov*	'string, thread'
	abstractions characterized	*sia*	'life'
	by length; nominalizations		
	with *kev* 'path, way'	*kev mob*	'sickness'
lub	Bulky inanimates; some	*rooj*	'table'
	abstractions	*tsev*	'house'
daim	Flat objects	*pam*	'blanket'
		ntawv	'paper'
rab	Tools	*keeb*	'needle, pin'
		phom	'gun'

There is also a general plural classifier *cov,* on which there are complex restrictions (Ratliff 1991).

Classifiers may perform disambiguating functions. For example, in Mandarin (Randy LaPolla, personal communication) the choice of classifier can distinguish two different senses of a noun, as in (1) and (2).

(1) *yī dào mén*
one CLF door
'a doorway'

(2) *yī shàn mén*
one CLF door
'a door' (i.e., the slab)

Another type, illustrated in Hmong in (3), compounds a classifier and noun to create a new word, although the classifier still performs its usual grammatical functions (Ratliff 1991). *Tus* classifies certain long things.

(3) *tus dej*
CLF water
'river'

The grammatical distribution of classifiers is also subject to complex conditions. Classifiers are generally required in NPs with one or more of the following: quantifiers, interrogative 'which', and demonstratives. Examples (4)-(10) are from Thai, and (5) and (10) are anaphoric.

(4) *náŋsɤ̌y sǎam lêm*
book three CLF
'three books'

(5) *sǎam lêm*
three CLF
'three' (of the *lêm* class)

(6) *cháaŋ lǎay tua*
elephant several CLF
'several elephants'
(Conklin 1981: 71)

(7) *sôm kì baj*
orange how-many CLF
'how many oranges?'
(Brown 1968: 61)

(8) *rót khan nǎy*
car CLF which
'which car' (Brown 1968: 161)

(9) *dèk khon níi*
child CLF this
'this child'

(10) *khon níi*
CLF this
'this' (of a person)

In Hmong, a classifier is also obligatory in most possessive NPs, as in (11)-(12):

(11) *kuv lub tsho*
1SG CLF shirt
'my shirt'

(12) *kuv lub*
1SG CLF
'mine' (of the *lub* class)

There are also complex semantic and discourse conditions on the occurrence of bare NPs versus NPs of the form N + *one* + CLF in HMT (Conklin 1981; Riddle 1989b; Bisang 1993). Just a small sample is given here.

Mandarin may combine a noun and classifier to express a plural or collective sense (Li and Thompson, 1981: 82, 34), as in (13), and reduplicate the classifier to mean 'every' instance of the referent, as in (14):

(13) *shū běn*
book CLF
'books'

(14) *kē kē shù*
CLF CLF tree
'every tree'

In Hmong, classifier presence varies according to degree of specificity, givenness and discourse salience (Riddle 1989b), as shown in (15), a passage from a Hmong novel, where Laotian communists fight the French.

(15) . . . *Fab-kis xuas*
. . . French take
(a) ***dav-hlau*** *tuaj*
plane come
'. . . The French came with planes'
Koos-les xuas tuam-phom txais
Communists take gun shoot
'the Communists shot at [them]'
(b) ***Fab-kis*** **cov** ***dav-hlau*** *ntuj-ntwg muaj*
French PL.CLF plane thunder have
'the planes of the French thundered'
(c) ***ib lub dav-hlau*** *quaj*
one CLF plane scream
'one plane screamed'
Yaj tawm hauv tsev tuaj nrog nws txiv ntsia
Yaj leave in house come with his father watch
'Yaj came out of the house with his father and watched'
(d) ***lub dav-hlau*** *poob*
CLF plane fall
'the plane fell. . .'
(e) ***Lub dav-hlau*** *ncho-pa lug tshwm tom-qab*
CLF plane smoke much appear behind
'Smoke appeared at the plane's tail end'
thiab ya tawg loj tawg lees-qis
and fly break big break drift
'and it flew down, breaking into big pieces and drifting'
zuj-zus rau hauv cov hav thuv ntawm ntug Xeev
gradually to in PL.CLF grove pine at edge Xeng
'gradually into a pine grove at the edge of Xeng.'

ces nrov ib teg thiab ncho-pa dub-nciab tawm tuaj
then noise once and smoke black-black leave come
'then there was a noise and dark black smoke came out.'
. . . lawv tsis muaj leej-twg tau pom dua
they not have anyone ATT see again
'None of them had ever seen'

(f) ***dav-hlau*** *poob los li*
plane fall at all
'a plane fall (shot down) before.' (Lis 1986: 6; my translation)

The NP *dav-hlau* 'airplane' in (a) does not have the plural *cov* classifier, although the contextual interpretation is plural. Specific planes are referred to, but introduced existentially and not individuated. As Hopper and Thompson (1980) claim for such NPs in other languages, they are not referentially salient. What is important is the presence of the planes, not their individual characteristics.

NP (b) *Fab-kis cov dav-hlau,* a possessive plural meaning 'the planes of the French', is a second mention of the planes. This is an environment normally requiring a classifier, but a more basic motivation for the classifiers in possessive NPs is that the possessive makes the reference specific (Riddle 1989b). Here, the plural classifier *cov* highlights the importance of the planes in a way that a non-possessive lacking the classifier would fail to do.

NP (c) *ib lub dav-hlau* 'a/one plane' singles out a specific plane. NP (d) *lub dav-hlau* 'the plane' refers back to this plane and constitutes given information. If *lub* were omitted in (d), the NP would refer to one of the group referred to by *Fab-kis cov dav-hlau* 'the planes of the French' in (b), but not necessarily the same airplane as that in (c) *ib lub dav-hlau* 'a/one plane'. NP (e) *lub dav-hlau* 'the plane' refers to the same topical plane. Inclusion of the classifier here is necessary or strongly preferred, since the plane is highlighted with descriptive details.

Finally, (f) *dav-hlau* 'plane' is a general reference and does not refer to a specific plane. For this reason, no classifier or number one-plus-classifier is used. No individual plane, group of planes, or individual characteristics of planes are highlighted here.

Even taking into account the proliferation of forms in synthetic languages where gender, number, and case may be indicated by a single affix, HMT must be considered to have quite complex systems of noun classification and NP syntax. Compare the declension of the masculine noun *fílos* 'friend' in Greek in (16), illustrating the maximum number of phonologically different nominal case forms. (Holton et al. 1997: 51, my transliteration).

(16)

fílos	N.SG	*fíli*	N.PL, V.PL
fílu	G.SG	*fílon*	G.PL
fílo	A.SG	*fílus*	A.PL
file	V.SG		

Due to syncretism, the number of phonologically distinct cases pronounced is smaller than the number of grammatical cases. As Carstairs-McCarthy (1994) argues, however,

the inferences drawn from the use of a lexical item of a given declension class in a particular context supply grammatical information obscured by the homophony of some inflectional affixes. This could be considered a complication of the system. For example, the unmodified noun *fíli* will be interpreted either as a nominative plural or a vocative plural based on its function in a clause. Likewise, *fíli* is interpreted as a nominative plural versus a genitive singular, as in *klé́fti* 'thief' (also pronounced the same in the accusative) given its grammatical function and the forms of any determiners and modifiers, but also because the nominative singular ends in *-os* as opposed to the *-is* of *kleftis*. Clearly, its paradigmatic relationships provide an important type of grammatical information.

In short, the Greek case system could be considered simpler than the HMT classifier systems with respect to numbers of categories. In addition, the relational patterns in HMT which disambiguate noun reference or express specificity, givenness and salience via classifier inclusion are quite parallel to the relational information which disambiguates Greek case marking. As will be shown below, relational information derived from lexical juxtaposition is important for understanding verb serialization and aspect in HMT as well.

The above discussion has provided a brief description of some typical classifier behavior. I suggest that classifier systems contribute to lexical elaboration, a form of explicitness and complexity, in two ways. First, although classifiers have characteristics of nouns, and/or quantifiers, they are best considered a separate word class, based on their grammatical distribution. Second, the use of classifiers as opposed to bound declensional affixes adds a semantically rich lexical item to many types of NPs, sometimes performing a disambiguation function.

Given the arbitrariness of human language, there is little independent justification for considering bound inflections to be more complex than a classifier system. (Even the morphophonological complexity of inflection is not an issue, since phonological processes can operate across word boundaries as well. This is especially true of tone sandhi in HMT. See e.g., Ratliff 1992a, b.) A useful comparison is to the use of prepositions as case markers in English versus the bound case markers of Greek and Polish. From an information-theoretic stance, the English system using primarily independent words as case markers would apparently be less complex than the Greek and Polish systems. However, in terms of ease of foreign language acquisition, considered by Kusters (this volume) to be a complexity metric, the comparatively large system of English prepositions would be quite complex. In fact, although English speakers may find the Polish case system daunting, according to Covitt (1976, cited by Celce-Murcia and Larsen-Freeman 1983: 340), English prepositions are the second most difficult aspect of English for foreign language learners, the first being the article system, also involving separate words as grammatical markers.

Compare the English article system to NP inflection in Polish, a language without articles. It seems fairly easy from a descriptive stance to list the possible inflections of Polish nouns, adjectives and determiners and their grammatical functions, despite the three-gender system, case differences and special features. As a native English speaker

who learned Polish while working in Poland, I did find the noun genders and inflections a memory burden, but not conceptually difficult. This raises the question of whether it is objectively possible to compare the complexity of different cognitive processes without experimental evidence.[3] In any case, linguists have long struggled to accurately represent the English article system (Hawkins 1978; Chesterman 1991; Lyons 1999; Epstein 2001), and a lengthy description is necessary, including information impinging on article presence versus absence, definite versus indefinite, and article pronunciations. The interpretation of a noun as mass or count in a given context, and if count, its number, must be considered, along with uniqueness or inclusiveness versus exclusiveness (Chesterman 1991), and locatability in a mental set (Hawkins 1978; Chesterman 1991) or accessibility in a mental space (Epstein 2001), depending on the analysis. In fact, in my 30 years of experience teaching adult native speakers of languages lacking articles, I have found that many such learners continue for many years not only to make mistakes with articles, but also to find them puzzling.

2.1.2 *Verb Serialization*

Another type of lexically elaborated structure is the serial verb construction. This is defined here as the concatenation of two or more verbs without overt marking of subordination or coordination. Intervening NPs may occur. A Thai example is given in (17) (Jagacinski 1992: 129, transcription, glosses modified).

(17) ***aw*** *khanŏm* ***pân*** *rɛw rɛw khâw . . .*
take dessert shape quick quick increasingly . . .
'Shape the (dessert) dough quickly. . .'

Dahl (2004) considers linearity to be a less complex device than inflection for indicating grammatical properties. This seems reasonable if only linearity *per se* is considered. But the juxtaposition of lexical items involves more than pure linearity: as each lexical item is added to the mix, sets of semantic properties accordingly are added, any of which might interact with another, resulting in complex semantic interrelationships in the construction of meaning. For example, while the verb in (18) has an activity sense, the substitution of a direct object for the adverb results in an accomplishment reading (Vendler 1967, my example).

(18) *I wrote for an hour.*

(19) *I wrote a letter.*

3. Although linguists assuming an autonomous syntax in a modular approach would separate this general cognitive issue from the study of grammar, other linguists would consider them to be interrelated.

Likewise, in HMT, not only are there restrictions on the relative linear ordering of serial verbs, but also rules governing the obligatory versus optional presence of particular verb types to express specific grammatical functions as well as lexical meanings. A Hmong example is given in (20), where the presence of either *los* 'come' or *mus* 'go' is obligatory (Riddle 1989a: 5, based on Li et al. 1986).

(20) *Lawv* ***nce*** ***nkag*** ***los/mus***.
3PL climb enter come/go
'They climbed in.'

In addition to expressing locomotion, *los* 'come' and *mus* 'go' indicate movement toward and away from the speaker, respectively, a distinction lacking in the corresponding English sentence.

Jarkey (1991) claims that Hmong is sometimes lexically less complex than English in that smaller meanings may be encoded in single lexical items. In her view, this may motivate serialization to express what might be encoded in fewer verbs in a language such as English.

Given Dahl's (2004) analysis of meaning compression as complex, some individual HMT verbs might arguably be less complex than some verbs in non-serializing languages. However, there are three problems with this possible view of HMT:

1. There are other HMT verbs, not to mention other types of lexical items, which would be more complex by the same criterion.
2. As in (20) above, many serial verbs have abstract grammaticalized functions involving deixis, aspect, or mood. (Li and Thompson 1981; Li, et al. 1986; Thepkanjana 1986; Riddle 1989a; Jarkey 1991)
3. Even if there is less meaning compression in HMT serial verbs, I argue that this is counterbalanced by the fact that serialization patterns are highly constrained, which contributes a different type of complexity.

Elsewhere, I have argued (Riddle 1990) that serialization and other forms of verbal concatenation, whether or not they are defined as serialization proper, also serve an elaboration function by providing a richness of detail in lexically referring to subparts of an overall situation with a string of verbs. In comparison, a language such as English often leaves those subparts unsaid. A Hmong example is given in (21):

(21) . . . *zoo-li* *yog* *nkawd* ***sawv*** ***nres*** ***tos*** *bus*
as be 3DUAL stand stop wait bus
'. . . as though they were waiting for the bus.' (Chang 1986: 5)

In (21), *tos* 'wait' can also occur without the other, underlined verbs, and *sawv* 'stand' plus *tos* can occur without *nres* 'stop'. The effect of including all three verbs is to make explicit that the two people seen by the speaker are stationary. The serialization elaborates explicitly details of a situation which might merely be inferred in Greek, Polish and English. Extending Gil's (this volume) analysis of associational meaning,

the fact that such meanings would be available only inferentially in Greek, Polish and English would make those languages less complex in this regard.

Another type of serialization explicitly represents with a second verb the result of the action indicated by a preceding verb, as in (22) from Hmong.

(22) *Nws* ***nyeem*** *ntawv rau kuv niam* ***nloog***
3SG read book to my mother listen
's/he's reading to my mother' (Strecker and Vang 1986: 14)

In (22), although it is reasonable to infer that the mother listens when read to, and that therefore *nloog* 'listen' is redundant, in Hmong explicit reference to the result is necessary here. Such resultative constructions are extremely common in HMT and other highly analytic SEA languages (Li and Thompson 1981; Zhang 1991; Matisoff 1973). In terms of Dahl's (2004: 54) notion of cross-linguistic dispensability, the Hmong structure expressing result in (22) would seem to be less complex than the English translation equivalent, since the use of 'to' in English is adequate without overt reference to listening. However, this means that the listening is left to be inferred, which seems comparable to the role of inferencing syntactic or semantic (see Gil, this volume) relationships from juxtaposition in highly analytic languages like HMT. Thus, whether one is describing the semantics of relatively synthetic or analytic languages, inferencing relative to the absence of some linguistic element, be it lexical or morphological, occurs. Either way, redundancy and conventional patterning are involved.

2.1.3 *Compounding*

Compounding is the most common word formation process in highly analytic languages such as HMT. As noted above, from the purely formal point of view, HMT-style compounds of two or more otherwise independent morphemes represent a different typology from derivation through affixation. They are more formally linear than inflectional morphemes since they usually simply juxtapose two free morphemes, although tone sandhi sometimes occurs. However, a very common phenomenon in HMT is the compounding of two independent morphemes which are either near synonyms or otherwise members of the same semantic field. Such a compound is an elaborated alternative to a single morpheme word with the same basic meaning as one member of the compound, although there may be some slight sense or stylistic difference, depending on the particular case. Some examples from Mandarin are given in (23)–(24) below.

(23) *jīn* 'gold' + *qiǎn* 'money' > *jīn qiǎn* 'money'

(24) *shuì* 'sleep' + *jiaò* 'fall asleep' > *shuì jiaò* 'sleep' (Li & Thompson 1981: 74)

This type of compounding is extremely productive in HMT.

2.1.4 *Elaborate expressions*

Related to the compounding predilection described above is the existence in HMT of lexicalized and productively derived sequences of four monosyllabic morphemes called

"elaborate expressions" (term from Haas 1964: xvii). Depending on the language, there are different sub-patterns for the meaning and form properties of the syllables. For example, in both Hmong (Johns and Strecker 1987) and Thai (Haas 1964), usually the first and third syllables match in some way, as do the second and fourth syllables. This may comprise full reduplication, or phonetic and/or semantic similarity. In Mandarin (Li and Thompson 1981), all four syllables may be different in form and meaning, but often two of the four may match in some way, or there may be two matching sets. In all three languages, elaborate expressions range from colloquial to elegant. Sometimes, the elaborate expression is the conventional lexical item for a particular concept, as in the Hmong word for 'storm' in (25).

(25) *cua daj cua dub*
wind yellow wind black
'storm' (Johns and Strecker 1987: 109)

Other elaborate expressions are relatively more descriptive alternatives to ordinary words or phrases, similar to the two-part compounds discussed above. Matisoff (1986: 76, 1992: 190) claims that both these phenomena are typical of the highly analytic languages of the region, and represent a tendency to "bulk up", thus counteracting a tendency toward monosyllabicity. Although in some cases, elaborations could be motivated by a need to disambiguate homophonous words by adding further description, this does not seem to be the case in general.

Other Hmong examples are given in (26)–(27). Note that in (26), *leej* is a marked classifier restricted to humans and *tus* is the usual classifier for humans, as well as for animals and long, firm objects.

(26) *Coob leej ntau tus*
numerous CLF many CLF
'many people' (Dao 1987: 5)

(27) *nroo tshaib nroo nqhis*
sigh hunger sigh thirst
'starve' (Tus txheej txheem hauv Indochina 1987: 46)

In Hmong, it is also possible to split compounds, as in (28).

(28) *poob teb poob chaw*[4] (cf. *teb chaws* 'country')
fall land fall place
'leave one's country' (*Dej cawv mob*, n.d.: 4)

4. Tone sandhi here changes *chaws* to *chaw* after the "b" tone.

Elaborate expressions are extremely common in HMT discourse, with the particulars concerning register varying across items and languages. They are yet another example of how HMT typology favours lexical elaboration as opposed to morphologically-based grammatical elaboration. I argue that these are very parallel forms of complexity, although belonging to different areas of the grammars of highly analytic versus highly synthetic languages.

Some theorists might argue that such phenomena belong to the pragmatics of language use or style, and are therefore irrelevant to the calculation of the grammatical complexity. However, there are constraints on the possible forms lexical elaboration takes in HMT, and these are not subject to cross-linguistic inferencing based on general knowledge. Moreover, according to Gil (this volume), style differences in Javanese, at least (based on interlocutor social status), do contribute to the complexity of Javanese as opposed to Riau Indonesian. Significantly, knowledge of elaborate expressions is fundamental to "sounding native" in HMT. According to my consultants, the spoken Hmong of many young Hmong-Americans (for whom Hmong is their first language) is often criticized by older speakers for the absence of various forms of lexical elaboration. This seems very parallel to loss of affixation, which is a hallmark of the development of pidgin and creole languages, and considered by McWhorter (1998, 2001a) and others to represent a decrease in complexity.

2.1.5 *Complexity in language learning*

In contrast to Dahl's complexity model, Kusters (this volume) argues for a complexity metric in terms of ease of learning as a foreign language. Lindström (this volume), however, argues that ease vs. difficulty of foreign language learning must be considered relative to a speaker. Therefore, the determination of linguistic complexity must be separated from difficulty of foreign language acquisition. In relation to these analyses, I also note that any discussion of difficulty must make clear what level of achievement in a foreign language is being taken as a norm.

The notion of what constitutes the "learning" of a foreign language is itself a major issue in the field of second language acquisition (Krashen 1981), and both practical and theoretical problems associated with testing degree of success or proficiency continue to be discussed.

It is instructive to look at data from the Foreign Service Institute (FSI) of the United States (Omaggio 1986: 21, citing Liskin-Gasparro 1982) on levels of speaking proficiency generally expected of native English speakers for four groupings of languages taught there in relation to three levels of aptitude and length of training. Significantly, Finnish, a highly synthetic language, is grouped with Vietnamese, a highly analytic language. These languages are both written in the same script as English and neither target language is Indo-European, unlike English, although both have Indo-European loanwords. According to the FSI scale, whether studying Finnish or Vietnamese, it takes a native English speaker of superior language learning aptitude some 1,300 hours of training to achieve level 3 speaking proficiency,

characterized by the ability to communicate readily and be understood by native speakers on most topics, but still making errors in various areas and having problems in style shifting.

Another significant comparison is that for the same type of learner, it takes some 720 hours of instruction to achieve the same level of proficiency in Haitian Creole, a relatively analytic language and Swahili, which is a synthetic language of the agglutinating type. Clearly, neither the presence nor absence of a significant degree of inflection accounts for these groupings.

Although at first glance, the lack of inflection in HMT might make it seem simpler for a language learner to construct a basic sentence in these three languages than in Greek, Polish and English, this is misleading, since there are many other sentence types which are quite complex in other ways. For example, the presence of so-called optional devices, such as aspect markers, indicators of specificity, and the like, is still subject to complex conditions on how they may or must be combined to express particular meanings.

For example, Haas (1964: xx) points out that the Thai sentence in (29) can be translated in four ways into English:

(29) *máa wîŋ rew*
horse run fast
a. 'The horse runs fast.'
b. 'The horses run fast.'
c. 'The horse ran fast.'
d. 'The horses ran fast.'

Generic (horses) and progressive (be running) interpretations are possible as well. This might make it seem easier for English speakers to learn Thai than for Thai speakers to learn English, and that therefore, Thai is grammatically simpler than English in this respect. However, I contend that this relative simplicity is more apparent than real. In fact, the proposed readings of the bare-bones Thai sentence are not equally natural, and it is much more likely to be used generically than in the other senses. This has partly to do with the fact that pragmatically, horses are generally expected to run fast. Although this encyclopedic knowledge is not part of the grammar of Thai, knowledge of the conventional ways to express relationships between particular forms and meanings is. Moreover, the form in (29) cannot be used in every situation where the bare proposition holds truth conditionally, since Thai marks other types of meaning not represented in the English translations above, making the apparent ambiguity of (29) irrelevant. Thus, although Thai does not mark past tense, for example, it has quite a few aspect markers, any of which, if included in (29), would express distinctions not made in English. For example, the grammaticized marker *lέεw* expresses a particular type of perfectiveness, reference to one or more boundaries (cf. Chiravate 2002, who describes this in terms of "abutment"). In fact, *lέεw* is required when there has been a clear change of state, as in (30) (my data), taken from a magazine article (Let it B 2000: 60) about a Thai movie star famous for being single.

(30) *. . . saaw bii tɛŋŋaan sĭa lɛ́ɛw*
miss Bii marry break PERF
"Miss Bii has (unexpectedly) gotten married."

The word *sĭa*, for the sake of brevity glossed literally here as the verb 'break,' shows that an extreme and unexpected boundary has been crossed. In this context, *lɛ́ɛw* is required since it is a boundary marker. Its presence is thus determined by the intended meaning of the sentence, indicated by the co-occurrence of *sĭa*. Hmong and Mandarin also have highly constrained perfective markers of a similar sort, not to mention other aspectual distinctions as well. According to Li (n.d.), a major problem in the description of aspect in (South)east Asian languages is some linguists' tendency to study isolated sentences, and I would add, in translation. This results in a failure to consider crucial conditioning factors on the occurrence of aspect marking, which is not simply optionally emphatic, contrary to popular belief.

Regarding ease of learning a particular language, a comparison may be made to playing the violin versus the piano. Although I doubt that anyone knowledgeable in music would claim that the task of a professional violin soloist is more difficult than that of a pianist, it is nonetheless true that for a beginner, it is easier to play a given note on the piano, where each note is associated with a separate key, than on the violin, where the strings form a continuum and the player must draw the line. This does not mean, however, that it is easier to become an Emanuel Ax than an Itzhak Perlman.

Why, then, do languages in contact situations tend to minimize inflection? As McWhorter (1998: 793) observes,

> . . . rapid non-native adoption of a language as a lingua franca entails stripping down a system to its essentials, for optimal learnability and processibility. The natural result is the virtual or complete elimination of affixes, sometimes replaced by more immediately transparent analytic constructions.

In other words, *initially*, inflected forms *are* more difficult than bare forms to learn and use consistently, since the forms of inflectional morphemes are completely arbitrary and highly language-specific. While some grammatical relationships may be amenable to calculation, inflections must be specifically memorized. This would appear to place a high price on inflectional morphology for second/foreign language learners. Nonetheless, as a language learner moves into more natural communication at higher levels of competence, other forms of complexity may show up which are not apparent to the beginner.

3. Conclusion

I have argued from two standpoints that highly analytic languages such as HMT are as grammatically complex as more synthetic languages. Although they eschew overt paradigmatic grammatical markers such as bound inflections, they favour elaboration on the lexical level, a syntagmatic property of the grammar. First, HMT have complex

subsystems of classifiers and aspect marking, which in the very least, correspond in complexity to nominal gender/case marking and verb tense marking, respectively, in synthetic languages. Second, and more broadly, HMT have a very strong tendency toward lexical elaboration in multiple ways, which can be considered a typological characteristic analogous to the elaboration of surface grammatical marking in inflected languages. I have refrained from claiming that any particular feature is specifically compensatory, however, since, although compensation may occur in some cases in some languages, this notion seems to rely on the conception of one form as more "normal" than another, or one as the point of departure. Rather, I take the emic view that each language has what it has independently. Although cause and effect relationships may exist historically within a given language, as for example, the use of tones in HMT following the loss of a number of final consonants, this is different from considering the ensuing change to be teleological in preserving complexity per se, as opposed to preserving relevant distinctiveness. I suspect that a "tipping point" model (Gladwell 2000) somewhat parallel to the results of the accumulation and compounding of evolutionary biological changes is a better model of language change.

A crucial issue which arises is what the "grammar" of a language is considered to be in the first place. Many linguists, especially in the Chomskyan tradition, consider contextual information to be an extralinguistic feature of language use, rather than determinant of grammatical structure. Yet, as discussed above, in order to utter native-like sentences regarding the swiftness of horses in Thai, for example, a number of complex decisions must be made, such as the need to include in a given context and for a given propositional meaning any of several different aspect markers, time adverbials, quantifier plus classifier, classifier plus demonstrative, among other choices, all with specific linear ordering constraints and co-occurrence restrictions based on the possible oppositions and conventional meanings of particular constructions given particular lexical items in specific linguistic contexts. Examination of such phenomena in a corpus of connected HMT speech will show that these are not merely free choices, like flavours of ice cream. Rather, their occurrence is rule-governed and associated with subtle differences in interpretation, which may be different from those made in other languages, and which are not merely forms of emphasis.

It seems that one must redefine grammar in an ad hoc way in order to privilege surface marking over relational and individual lexical information. Thus, whether relations are overtly marked by affixation as opposed to other means is not relevant to judging the overall complexity of the grammar of a language. It would be more insightful, I suggest, to reframe the complexity discussion in terms of the degree to which a grammar displays regularity and exceptions, and the historical and typological factors involved therein, rather than to focus on complexity *per se*.

Abbreviations

1	first person
3	third person

A	accusative
ATT	attainment aspect
CLF	classifier
G	genitive
HMT	Hmong, Mandarin and Thai
N	nominative
PERF	perfective
PL	plural
SG	singular
V	vocative

References

Benedict, P.K. 1975. *Austro-Thai Language and Culture with a Glossary of Roots*. New Haven: Human Relations Area Files Press.

Bisang, W. 1993. Classifiers, quantifiers, and class nouns in Hmong. *Studies in Language* 17(1): 1–51.

Bisang, W. 1996. Areal typology and grammaticalization: Processes of grammaticalization based on nouns and verbs in East and Southeast Asian languages. *Studies in Language* 20(3): 519–597.

Brown, J.M. 1968. *AUA Language Center Thai Course. Book 2.* Bangkok: American University Alumni Association Language Center.

Carstairs-McCarthy, A. 1994. Inflection classes, gender and the principle of contrast. *Language* 70(4): 737–788.

Celce-Murcia, M. & Larsen-Freeman, D. 1983. *The Grammar Book*. Rowley MA: Heinle & Heinle.

Chesterman, A. 1991. *On Definiteness: A Study with Special Reference to English and Finnish*. Cambridge: Cambridge University Press.

Chiravate, B. 2002. Syntactic and Semantic properties of Aspectual Markers in Thai. PhD Dissertation, Michigan State University.

Conklin, N.F. 1981. The Semantics and Syntax of Numeral Classification in Tai and Austronesian. (Volumes I and II.) PhD Dissertation, University of Michigan.

Covitt, R. 1976. Some problematic grammar areas for ESL teachers. MA Thesis. UCLA.

Dahl, Ö. 2004. *The Growth and Maintenance of Linguistic Complexity* [Studies in Language companions series 71]. Amsterdam: John Benjamins.

DeGraff, M. 2001. On the origin of creoles: A Cartesian critique of neo-Darwinian linguistics. *Linguistic Typology* 5(2–3): 213–310.

Epstein, R. 2001. The definite article, accessibility, and the construction of discourse referents. *Cognitive Linguistics* 12(4): 333–378.

Foley, W.A. & Olson, M. 1985. Clausehood and verb serialization. In *Grammar Inside and Outside the Clause*, J. Nichols & A.C. Woodbury (eds), 17–60. Cambridge: Cambridge University Press.

Gil, D. 2001. Creoles, Complexity and Riau Indonesian. *Linguistic Typology* 5(2–3): 325–371.

Gladwell, M. 2000. *The Tipping Point: How Little Things Can Make a Big Difference*. New York: Little, Brown & Company.

Gordon, R.G., Jr. (ed.), 2005. *Ethnologue: Languages of the World*, 15th ed. Dallas, Tex.: SIL International. [Online version: <http://www.ethnologue.com/>].

Haas, M. 1964. *Thai-English Student's Dictionary*. Stanford: Stanford University Press.

Hawkins, J. 1978. *Definiteness and Indefiniteness: A Study in Reference and Grammaticality Prediction*. London: Croom Helm.

Heimbach, E.E. 1969. *White Hmong-English Dictionary*. Ithaca: Southeast Asia Program, Cornell University.

Holton, D., Mackridge, P. & Philippaki-Warburton, I. 1997. *Greek: A Comprehensive Grammar of the Modern Language*. London: Routledge.

Hopper, P.J. and Thompson, S.A. 1980. Transitivity in grammar and discourse. *Language* 56(2): 251–299.

Hyman, L.M. 2000. How to become a "Kwa" verb. Paper given at the Symposium on areal typology of West African languages, Leipzig.

Jagacinski, N. 1992. The ʔau usages in Thai. In *Papers on Tai Languages, Linguistics and Literatures: In Honor of William J. Gedney on His 77th Birthday*, C.J. Compton & J.F. Hartmann (eds), 118–138. DeKalb IL: Center for Southeast Asian Studies, Northern Illinois University.

Jarkey, N. 1991. Serial verbs in White Hmong: A functional approach. PhD Dissertation, University of Sydney.

Johns, B. & Strecker, D. 1987. Lexical and phonological sources of Hmong elaborate expressions. *Linguistics of the Tibeto-Burman Area* 10(2): 106–112.

Krashen, S. 1981. *Second Language Acquisition and Second Language Learning*. Oxford: Pergamon Press Inc.

Li, C. N.d. The aspectual system of Hmong. Ms.

Li, C., Harriehausen, B., & Litton, D. 1986. Iconicity: A view from Green Hmong serial verbs. Paper given at the Conference on Southeast Asia as a Linguistic Area.

Li, C. & Thompson, S.A. 1981. *Mandarin Chinese: A Functional Reference Grammar.* Berkeley: University of California Press.

Liskin-Gasparro, J.E. 1982. *ETS Oral Proficiency Testing Manual*. Princeton, NJ: Educational Testing Service.

Lyons, C. 1999. *Definiteness*. Cambridge: Cambridge University Press.

Lyovin, A.V. 1997. *An Introduction to the Languages of the World.* Oxford: Oxford University Press.

Matisoff, J.A. 1973. *The Grammar of Lahu*. Berkeley: University of California Press.

Matisoff, J.A. 1986 [1983]. Linguistic diversity and language contact. In *Highlanders of Thailand*, J. McKinnon & W. Bhruksasri (eds), 56–86. Singapore: Oxford University Press. [Paperback 1986, originally published in hardback 1983].

Matisoff, J.A. 1992. The Lahu people and their language. In *Minority Cultures of Laos: Kammu, Lua', Lahu, Hmong and Iu-Mien*, J. Lewis (ed.), 145–247. Rancho Cordova, CA: Southeast Asia Community Resource Center.

McWhorter, J.H. 1998. Identifying the creole prototype: Vindicating a typological class. *Language* 74(4): 788–818.

McWhorter, J.H. 2001a. The world's simplest grammars are creole grammars. *Linguistic Typology* 5(2–3): 125–166.

McWhorter, J.H. 2001b. What people ask David Gil and why: Rejoinder to the replies. *Linguistic Typology* 5(2–3): 388–412.

Mufwene, S. 1986. Les langues créoles peuvent-elles être définies sans allusion à leur histoire? *Etudes Créoles* 9: 135–150

Omaggio, A.C. 1986. *Teaching Language in Context*. Boston: Heinle & Heinle.

Ramsey, S.R. 1987. *The Languages of China*. Princeton, NJ: Princeton University Press.

Ratliff, M. 1991. Cov, the underspecified noun, and syntactic flexibility in Hmong. *Journal of the American Oriental Society* 111(4): 694–703.

Ratliff, M. 1992a. *Meaningful Tone: A Study of Tonal Morphology in Compounds, Form Classes, and Expressive Phrases in White Hmong*. DeKalb, IL: Center for Southeast Asian Studies, Northern Illinois University.

Ratliff, M. 1992b. Tone language type change in Africa and Asia: !Xũ, Gokana, and Mpi. *Diachronica* IX(2): 239–257.

Riddle, E.M. 1989a. Serial verbs and propositions in White Hmong. *Linguistics of the Tibeto-Burman Area* 12(2): 1–13.

Riddle, E.M. 1989b. White Hmong noun classifiers and referential salience. Paper given at the XXII International Conference on Sino-Tibetan Languages and Linguistics, Honolulu, October 1989.

Riddle, E.M. 1990. Parataxis in White Hmong. In *When Verbs Collide: Papers from the 1990 Ohio State Mini-Conference on Serial Verbs* [Working Papers in Linguistics No 39]. B.D. Joseph & A.M. Zwicky (eds). Columbus: Dept. of Linguistics, The Ohio State University.

Riddle, E.M. 1992. Lexical elaboration in White Hmong. Paper given at the XXV International Conference on Sino-Tibetan Languages and Linguistics, Berkeley, October 1992.

Riddle, E.M. & Stahlke, H. 1992. Linguistic typology and sinospheric languages. In *Papers from the First Annual Meeting of the Southeast Asian Linguistics Society*. M. Ratliff & E. Schiller (eds), 351–366. Tempe: Program for Southeast Asian Studies, Arizona State University.

Sapir, E. 1921. *Language: An Introduction to the Study of Speech.* New York: Harcourt, Brace & World.

Strecker, D. and Vang, L. 1986. *White Hmong Dialogues*. Minneapolis: Southeast Asian Refugee Project, University of Minnesota.

Thepkanjana, K. 1986. Serial verb constructions in Thai. PhD Dissertation. University of Michigan.

Vendler, Zeno. 1967. Verbs and times. In *Linguistics in Philosophy*. Ithaca NY: Cornell University Press.

Zhang, B. 1991. Serial verb constructions or verb compounds? A prototype approach to resultative verb constructions in Mandarin Chinese. PhD Dissertation, Ball State University.

Corpus references

Chang, C. 1986. Hlub niam laus yuav niam hluas [Love the older sister, marry the younger sister]. *Haiv Hmoob* 2(1): 5–15.

Dao, Y. 1987. Liv Xwm Hmoob [Hmong History]. *Haiv Hmoob* 3(2): 5.

Dej cawv mob. (Alcoholism) N.d. Minneapolis: Community University Health Care Center.

Let it B. 2000. *LIPS*, March 2000, 60–84.

Lis, N. 1986. *Lub neej daitaw* [Suspended life]. Bonnyrigg, Australia: Roojntawv Neejmhoob.

Tus txheej txheem hauv Indochina [Reporter in Indochina]. 1987. *Haiv Hmoob* 3(2): 46–47.

Grammatical resources and linguistic complexity

Sirionó as a language without NP coordination

Östen Dahl
Stockholm University

The paper discusses the relationship between cross-linguistic differences in grammatical resources and linguistic complexity. It is claimed that Sirionó (Tupí-Guaraní) lacks syntactic coordination as in English *John and Mary are asleep*. Instead, Sirionó employs a number of different strategies – the 'with' strategy, the list strategy, and the 'also' strategy – to make up for this. It is argued that one or more of these strategies may serve as a diachronic source of syntactic coordination. The development of syntactic coordination in a language exemplifies condensation processes in grammaticalization and increases complexity in the sense that a certain type of complex syntactic structure is introduced, and makes it possible to express in one syntactic unit what previously needed two or more.

1. Background notions: System complexity, structural complexity, resources vs. regulations

Dahl (2004) is a monograph devoted to the notion of complexity and its applications in linguistics, in which complexity is seen as an absolute property of an object rather than a relational one defined in terms of the difficulty for a user. The basic idea is that the complexity of an object can be identified with the length of its shortest possible complete description or specification. This way of understanding complexity is not without its problems (some of them are discussed in my book), but will do as a basis for the ensuing discussion in this paper.

In Dahl (2004), I make a number of distinctions pertaining to the notion of complexity as applied to language. One important such distinction is between system complexity, which is a property of the language, and structural complexity, which is a property of expressions in a language. A couple of simple examples will illustrate these notions. In written English, the form of the definite article is always the same; in French, it depends on the number and gender of the following noun; in German, it also depends on case. In this respect (but not necessarily in others), German is more complex than French, and French is more complex than English. This comparison

pertains to the language system, so we are speaking of system complexity. The sentence *The man in the black coat opened the door to the kitchen* has a more complex syntactic structure than *He opened it*. This comparison, then, concerns structural complexity. Structural complexity is defined at the level of individual expressions, but it can also serve as the basis for global measures of texts and discourses, e.g., the average or maximal structural complexity of expressions that make up them.

Another fundamental distinction discussed in Dahl (2004) is that between "resources" and "regulations":

1. resources: what you can do, or: the set of possibilities that the system offers its users;
2. regulations: what you have to do, or: the constraints and requirements that the system puts on its users.

Speakers of a language have at their disposal a set of resources: an inventory of morphemes, words, phrases and constructions from which they can choose in building an utterance. But the process of building utterances also involves regulations, including but not restricted to what would traditionally be called "rules of grammar". For instance, in English, we may combine a noun phrase and an intransitive verb to form a sentence – a resource which enables us to express propositions that we want to communicate. But the resource is regulated by English grammar which demands that we put the noun phrase before the verb and that the verb should agree with the (subject) noun phrase. In Dahl (2004), I tried to show how diachronic processes such as those usually called "grammaticalization" can increase the complexity of linguistic regulations, mainly by introducing non-linearity into the expression. I did not have very much to say in the book about the possible application of complexity considerations to the resource side of language, in particular to the question what this would mean for a theory of language change: are there differences in complexity between languages that depend on one language having more resources than the other? And can a language become more complex by acquiring additional resources?

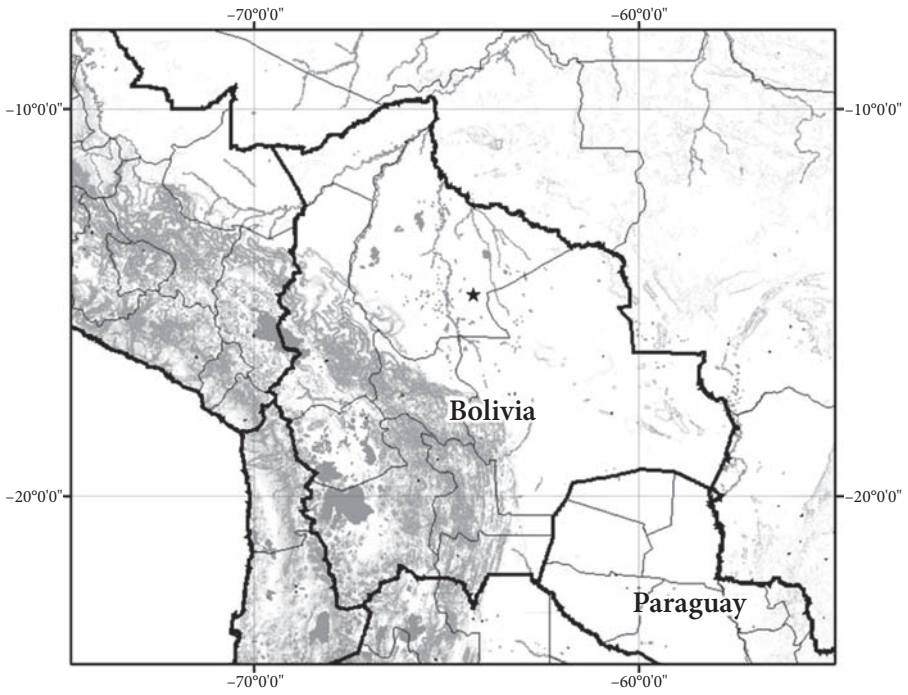
The simplest case of an addition of new resources is when new words are added to the lexicon of a language. As long as this just means that the number of lexical items increases, the effect on the system complexity of the language is somewhat trivial, and it may seem more natural to characterize a language with a large vocabulary as "rich" or "expressive" rather than "complex", although it is not obvious that having more words at one's disposal automatically makes it possible to express more or say things in more nuanced ways. And in fact, the choice between words is not always free, but governed by various kinds of regulations, which, arguably, influence system complexity in a more significant way. It is well known that grammars of languages force certain choices upon speakers whether they like it or not, but so does the lexicon: in Russian, one cannot say 'I came yesterday' without revealing both one's gender (by the choice of verb form) and one's mode of transportation (by the choice of verb lexeme). On these points, Russian is arguably more complex than English, while English is more complex

by requiring that speakers choose between the two lexical items *hand* and *arm* which correspond to the single word *ruka* in Russian. (Such lexical constraints exemplify regulations that are not easily formulated as rules in the traditional sense.)

If we move from lexical items to constructions, whether phraseological or grammatical, questions about complexity become in themselves more complex since grammatical structure enters the picture. The rest of the paper will be devoted to the discussion of a concrete example of a difference in grammatical resources between languages. The example concerns coordination, or rather the lack thereof, in Sirionó.

2. Sirionó, a language without syntactic coordination

Sirionó is an endangered Tupí-Guaraní language spoken by less than 1000 persons in the province of Beni, Bolivia, mainly in the village of Ibiato (Eviato, see map). The Sirionó were until quite recently hunter-gatherers.



Map. Location of Sirionó.

Sirionó, then, is a language which, as I will argue, lacks syntactic coordination. This means more concretely that there is no syntactic construction in the language which does what coordination (supposedly) does in a language such as English – that is,

takes as input two or more expressions of some category C and gives as output another expression of category C, which contains the input expressions as constituents of equal rank (that is, none of them is head).[1] Even more concretely, this means for instance that there is no exact equivalent to an English sentence such as *John and Mary sing*. Indeed, there is no word in Sirionó that can be said to translate *and*. The Spanish conjunction *y* 'and' is sometimes – and probably increasingly often – used but very rarely to conjoin noun phrases, but this is clearly not a feature of Sirionó as traditionally spoken. In the following, I shall restrict myself to noun phrase coordination, the reason being that for other types of coordination, e.g., of sentences, it is more often than not difficult to distinguish mere juxtaposition in the discourse from syntactic coordination proper.

What we see in Sirionó is possibly an areal feature of Amazonian languages. Derbyshire and Pullum (1986: 19) say that absence of coordinating conjunctions, "hence reliance on juxtaposition to express logical coordination" is a common feature of Amazonian languages. The following would be a fairly clear example of this from Dení (Arawakan, Derbyshire (1986: 538)):

(1) *u-kha da'u u-kha tu mede 0-vada-ru*
1-POSS son 1-POSS daughter 3PL 3PL-sleep-NONFUT.F
'My son and daughter are sleeping.'

In (1), however, there is still something that looks like a coordinated NP which is joined to a verb phrase as its subject. My claim is that Sirionó goes further in not having any counterpart to this syntactic structure. There may well be other languages in the area that are similar to Sirionó here, but many grammars of Amazonian languages (e.g., Dixon 2004: 528; Koehn & Koehn 1986: 52; Abbott 1991: 52) are rather vague on this point: even if they say that there is no word for 'and' in the language, they do not give any information what the speakers do if they want to say something like (1).

There is a general methodological problem here. If someone wants to show that a language lacks a certain construction, it may not be convincing enough to show that it is absent from recorded or written material; the reason may simply be that the speakers did not have any reason to use the constructions in those texts. Translated texts offer a way out in the sense that a systematic avoidance of a construction of a certain type, even when the source text would induce the use of such a construction, can be taken

1. Stassen (2000: 4) defines "NP-conjunction" more in terms of semantics, by postulating that a sentence containing a case of NP-conjunction should describe a single occurrence of an event (action, state, process, etc.) which is predicated of two or more referents conceived of as separate individuals. This definition is problematic in that it theoretically allows for any kind of syntactic configuration that has this semantics and also in that it is not always clear what a "single occurrence" means, for instance in a sentence such as *My brother and sister are students* (translation of one of Stassen's Russian examples (Stassen (2000: 38)).

as an indication of the absence of the construction in the language. Not surprisingly, the longest text that has been translated into Sirionó is the Bible, and given that the original Greek New Testament contains 9,268 occurrences of the word *kai* 'and', there should be ample opportunity to study what the translators have done when the original text contains a conjoined structure, even if only part of the occurrences involve NP coordination.

The Sirionó Bible translation exemplifies a number of ways in which it is possible to manage without syntactic coordination.[2] To start with a simple example, consider the translation of the beginning of Matthew 1: 11, which in the King James Version (KJV) reads *And Josias begat Jechonias and his brethren . . .* and in the Greek original *Iōsias de egennēsen ton Iekhonian kai tous adelphous autou . . .*:

(2) *Josías rei Jeconías ru. Jeconías nongue abe ru.*
Josias COP Jechonias father Jechonias brother also father
'Josias was the father of Jechonias. Jechonias' brothers' father, too.'

There is no syntactic unit in (2) that corresponds to the English coordinated noun phrase *Jechonias and his brethren* or the Greek *ton Iekhonian kai tous adelphous autou*. Rather, the effect of the coordinating construction is obtained by adding the phrase with the meaning 'Jechonias' brothers' father, too' as an after-thought, as it were. This is one of several strategies employed in the Sirionó New Testament to render coordinated constructions of European languages. I shall now discuss each of the strategies in detail, but let me first point out one general peculiarity of the Sirionó New Testament which (2) illustrates, namely that periods (full stops) occur more frequently than in English (or Greek, for that matter). That is, sentences tend to be split up orthographically in smaller fragments. Presumably, this is intended to reflect something like "intonation units" in the spoken language. At least for the spoken language examples below, this seems to make sense, as we shall see.[3] What I am not sure about is whether spoken Sirionó is really differently segmented than other languages. Probably, if we really tried to write the way we speak, English texts would look more like the Sirionó New Testament than they do now.

2. The Bible was translated into Sirionó in the seventies by the SIL linguist Perry Priest and the native speaker Chiro Cuellar. I have only had access to the New Testament.

3. The spoken materials referred to in this paper were transcribed by native speakers from video recordings made by the film-makers Mats Brolin and Anna Cnattingius in connection with the documentary film "Let our songs live", and in connection with a visit that Mats Brolin and I made to Ibiato in 1999.

3. How to manage without syntactic coordination: Three strategies

3.1 The 'with' strategy

Comitative constructions ('X with Y') are used widely in many languages instead of English *and*-constructions, and are also historically often the source of the latter. In the Sirionó New Testament, the postposition *rese/ndese* 'with' does occur fairly often when the original text contains a conjoining construction. A typical example is the following:

(3) *Emo mose que* ***Pedro*** *nyoɨ quia Dios chuchúaa ra.* ***Juan*** *ndese.*
one time PST1[4] Peter went IPFV God house PST2 John with
'Now Peter and John were going up to the temple . . . ' (Acts 3:1)

Again, we see that the English sentence is broken up in the Sirionó text. In many languages (3) would be translated as 'Peter with John went up to the temple', preserving the integrity of the conjoined subject. However, the noun phrases that are conjoined in the English version do not form a constituent in the Sirionó example; indeed, they occur quite far from each other, separated by a full stop. Although it is hard to exclude that the *rese/ndese* phrase could sometimes be contiguous with the other NP, I have found no such examples in the Bible text or in transcriptions of spoken conversations. It must be concluded that the 'with'-strategy does not involve syntactic coordination in any reasonable sense. In fact, in this respect the Sirionó use of the 'with'-strategy may not differ from what is found in many other languages: according to Stassen (2000: 18) it is a characteristic of what he calls the "Comitative Strategy" that the two NPs involved "are morphosyntactically encoded as NPs of unequal structural rank", that they "are not part of the same constituent", and that they do not force dual or plural agreement on predicates or are subject to restrictions on extraction rules. On the other hand, Stassen (2000: 26) also points out that pure instances of "WITH-languages" are relatively rare, since there seems to be a general tendency for the "Comitative Strategy" to develop gradually in the direction of bona fide syntactic coordination: "WITH-languages do not have a Coordinate Strategy, but they would like to have one". Sirionó might be an extreme case here in that there seems to be no tendency for comitative constructions to grammaticalize into coordination. In addition, even in the case of the languages that "take the comitative encoding as the only available option" (Stassen 2000: 21), the examples provided by Stassen overwhelmingly have the "conjoined" noun phrases next to each other, whereas this is not the case in Sirionó, as we saw above. However, this may be partly due to how Stassen has chosen his examples from the sources. Thus, for Urubú-Kaapor, Stassen presents (4), taken from Kakamasu (1986: 349), but on the same

4. Sirionó has a discontinuous past marker *que*/*ngue* . . . *ra*/*nda*, which is here glossed as PST1 . . . PST2.

page Kakumasu says: "Quite often the comitative occurs at the end of the sentence, following the main predication, functioning to clarify or specify", giving some examples that look very similar to the Sirionó way of using the 'with'-strategy, such as (5):

(4) *Nasúi riki ihẽ namõ i-hon.*
Nasúi EMPH me COM 3-go
'Nasúi went with me.'

(5) *Ma'e ke Kaitã ta kekar a'e ta o-ho, Te'oru namõ.*
something OBJ Kaitã PL 3.hunt 3 PL 3-go Te'oru also
'Kaitã folks went hunting, Te'õru also.'

3.2 The 'also' strategy

The following examples from (6) to (8) further illustrate the 'also' strategy, which we saw above in (2):

(6) *Saludar ere rã* ***Prisca*** *je.* ***Aquila*** *abe je. Onesíforo chuchua jenda ja je abe.*
greet say IMP Prisca to Aquila also to Onesiphorus house being all to also
'(literal translation:) Greet Prisca. Aquila also. The household of Onesiphorus also.'
'(KJV:) Greet Prisca and Aquila, and the household of Onesiphorus.'
(2 Timothy 4: 21)

(7) *Ibei ra ua. Ibi abe ra ua no.*
Heaven FUT pass-away earth also FUT pass-away also
'Heaven and earth will pass away . . . ' (Mark 13: 31)

(8) *Emo jiri mose que judíos ngaẽ huee ra. Antioquía si. Iconio si abe no.*
one more time PST1 Jew come here PST2 Antioch from Iconium from also also
'But Jews came there from Antioch and Iconium.' (Acts 14: 19)

In the 'also' strategy, the conjuncts from the second onwards are added after the sentence (within full stops) together with *abe (no)* 'also'. Smaller or larger parts of the sentence may be repeated.

3.2 The list strategy

In the list strategy, the coordinated NPs are just juxtaposed. In the New Testament translations, they are written with full stops between them, as in (9):

(9) *Ngasẽ ngue Jerusalén nda. Nyoi que uchuchúaa cote ra. Ã ja que ra.*
come PST1 Jerusalem PST2 go PST1 REFL.house.LOC now PST2 this all PST1 PST2

Pedro. Juan. Jacobo. Andrés. Felipe. Tomás. Bartolomé.
Peter. John. James. Andrew. Philip. Thomas. Bartholomew.
Alfeo riiri. Jabo, ɨ ee. Simón el Cananista. Judas. Jacobo
Alpheus son James say him Simon (Span.) (Span.) Judas James
riirī ndei Judas.
son COP Judas
'(literal translation:) They came to Jerusalem. They went into the house. All these. Peter. John. James. Andrew. Philip. Thomas. Bartholomew. Alpheus son. James by name. Simon the Canaanist. Judas. Judas was the son of James.'
'(KJV:) And when they were come in, they went up into an upper room, where abode both Peter, and James, and John, and Andrew, Philip, and Thomas, Bartholomew, and Matthew, James the son of Alphaeus, and Simon Zelotes, and Judas the brother of James. (Acts 1: 13)

Often this strategy is mixed with others. The following (10) illustrates a mixture of the list strategy and the 'also' strategy:

(10) *Ã ngue nyoɨ erese ra. Eata rete. Sópater. Berea jenda.*
this PST1 came with.him PST2 much much Sopater Berea being
Aristarco. Segundo abe. Tesalónica jenda rei ũquɨ̃. Gayo.
Aristarcus Secundus also Thessalonica being COP this Gayo
Derbe jenda. Timoteo.
Derbe being Timothy
Tíquico. Trófimo. Asia jenda rei ũquɨ̃.
Tychicus Trophimus Asia being COP this
'(literal translation:) These came with him. . . . Sopater, being from Berea. Aristarchus. Secundus also, these being from Thessalonica. Gayo, being from Derbe. Timothy. Tychicus, Trophimus, these being from Asia.'
'(KJV:) He was accompanied by Sopater son of Pyrrhus from Berea, Aristarchus and Secundus from Thessalonica, Gaius from Derbe, Timothy also, and Tychicus and Trophimus from the province of Asia.' (Acts 20: 4)

Similar mixed examples are found in spoken material. The following combination of the list strategy and the 'also' strategy in (11) is from a discourse about the Sirionó belief that various animals originate in the sky and then fall down on earth:

(11) *nyebe sira riqui nguiã ngoɨ ngoɨ. nguiquiare ngoɨ. mbeɨ mbae abe no.*
therefore fish COP IPFVV fall fall crocodile fall snake this also also
conombe. quiriicha. quereregẽ. ũquɨ̃ ngoɨ ngoɨ já te chee.
turtle kind.of.turtle water turtle this fall fall all EMPH until
'Therefore the fish fall. The crocodile falls. This snake also. The turtle. Small turtles. Water turtles. Until all these things fall.'

In the following spoken example (12), the 'with' strategy and the 'also' strategy are mixed:

(12) *ũquɨ̃ ndese chõ aɨco nguiã ndua ndua. se-resenda ja rese.*
this about FOC COP.1SG IPFV think think my-companion all with

se-nongue abe no.
my-brother also also
'This is what I am thinking about. With all my companions. My brothers also.'

4. Discussion and conclusions

In the rendering of the spoken examples, full stops indicate the boundaries between what I perceive as intonation units. It appears that in general, the phrases that correspond to English conjuncts are pronounced as separate intonation units, a fact that is well in congruence with the Bible translators' practice of separating them by full stops. I claimed above that there is no syntactic coordination in Sirionó. We now see what this means in practice. In Sirionó, the content of an English noun phrase will typically be split up and distributed over constituents that may be rather far from each other and usually show up in different intonation units. Let us now consider the theoretical implications of this.

It is clear that at least some of the "strategies" found in Sirionó are possible also in languages such as English. Thus, instead of (13) one may say (14):

(13) *Peter and John are coming to the party*.

(14) *Peter is coming to the party. John too*.

English, like most other European languages, thus has (at least) two possibilities of expressing one and the same content, one of which is lacking in Sirionó. Furthermore, there is a possible historical link between 'and' constructions, as in (13), and 'also' constructions, as in (14), in that some words for 'and' are supposed to be etymologically derived from words meaning 'also'. Thus, the Swedish word *och* 'and' is homophonous with the (now obsolete) word *ock* 'also'; in other North Germanic languages (Icelandic, Norwegian, Dalecarlian) the form *og* is used in both meanings, whereas the cognates in other Germanic languages, such as German *auch*, have the meaning 'also' only, suggesting a path of development from 'also' to 'and' and thus the possibility of a grammaticalization path from constructions like that in (14) to constructions like that in (13). A closer comparison of these two sentences is therefore well motivated.

Sentences (13) and (14) are equivalent in the sense that they give exactly the same information about the world – they have the same truth-conditions. But they differ in their pragmatic or rhetorical structure. While (13) represents one single speech act, the two parts of (14) are potentially two different speech acts, which can be seen from the fact that they may well be uttered by different speakers. However, even if we have two speech acts in (14), they are not wholly independent of each other, in that the interpretation of the second relates crucially to that of the first: the words *John too* cannot normally be discourse-initial. More precisely, when interpreting *John too*, we abstract out from the first part of (14) whatever is said about Peter and try to apply that to John. But (13) and (14) may also be said to differ semantically, more precisely, in the way the truth-conditional information is built up. In (13), the subject NP defines a set A of two

members, Peter and John, and as this NP is joined with the predicate VP, we learn that the members of A have a certain property *p*, namely coming to the party. Metaphorically, Peter and John are added to the list of party guests in one swoop. In (14), we first put Peter on the list of guests, then John is added as an afterthought as it were.

Summing up the semantic and pragmatic differences between the 'and' construction in (13) and the 'also' construction in (14), we may say that while (13) represents a *compositional* mode of interpretation, (14) can be said to be *incremental*. This is paralleled in the syntax in that whereas the 'and' construction creates a complex noun phrase out of two simple ones, the two noun phrases in the 'also' construction remain linearly and hierarchically separate. Above, we noted that 'also' constructions are plausible historical sources of 'and' constructions, which would mean that grammaticalization processes may lead from an incremental to a compositional mode of interpretation and from linear to hierarchical syntactic relationships. This may be seen as exemplifying the process that were referred to in Dahl (2004) as "condensation" and by Givón (1979: 82–83) as "processes by which loose, paratactic, *pragmatic* discourse structures develop – over time – into tight, *grammaticalized* syntactic structures" (emphasis added).

But let us first consider the proposed distinction between the compositional and incremental modes of interpretation in some detail.

The basic idea of an incremental mode of interpretation can be described in the following way. We have some device that has a set of possible internal states, which receives a linear sequence of instructions, where each instruction specifies one operation by which the device passes from one state to another. Each operation then applies to the state that the device is in at that point in time. We can illustrate this idea by the following sequence of instructions (15) which can be executed on a pocket calculator:

(15) An incremental example:

Instruction	Internal state of the device
Enter 5	5
Add 7	12
Subtract 2	10
Multiply by 4	40

In a system with a compositional mode of interpretation, all these instructions can be squeezed into one complex expression:

(16) A compositional example:
4*(5+7–2)

The hierarchical structure of the expression, rather than the linear ordering, tells us to how to apply the instructions. This will become clearer if we may use a prefix notation instead of the traditional arithmetic one:

(17) MULTIPLY(4, SUBTRACT(ADD(5,7), 2)))

How does all this relate to the issue of complexity? It is clear that a compositional mode of interpretation involves expressions with a greater structural complexity than an incremental mode of interpretation. A pocket calculator which allows for expressions like (16) or (17) thus employs a "language" with a greater maximal structural complexity than one which only allows for incremental sequences like (15), and to the extent that the more complex expressions demand more complex descriptions, the language will also have a greater system complexity. But it is worth noting that the actual operations carried out by the calculator are not necessarily different in the two alternatives, and irrespective of the mode of interpretation, the calculator has to "remember" the result of the preceding operation, since that is what the next operation applies to. There is a clear parallel here to the linguistic examples (13) and (14). The interpretation of an expression like *John too* in (14) has to operate on what is already there in the listener's memory, that is, it has to connect to the interpretation of the preceding sentence, which may be seen as the creation of an anaphoric link. The use of the coordinate structure in (13) obviates this need for such a link, but at the price of a more complex syntactic structure at the sentence level.

So, to sum up, suppose that Sirionó borrows NP coordination from Spanish (something that may actually be happening). How does this affect complexity? First of all, it affects the complexity of the language in a similar way that the addition of a new word does – which was said above to be somewhat trivial. On the other hand, it can be argued that coordinating constructions of the kind that we know from other languages have certain quite specific and probably unique syntactic properties; it stands to reason that the introduction of such constructions will increase the overall complexity of grammar, i.e., system complexity. Looking at global properties of discourse, we can see that since the coordinating construction would be used in lieu of expressions with a simpler syntactic structure, it would also increase the average structural complexity of sentences, although some connections at the discourse level may simultaneously become superfluous. However, since Sirionó already contains other constructions of a complexity which is comparable to that of coordinating constructions, it is not obvious that it also influences the maximal structural complexity of constructions. We can thus see that the ways in which the increase in grammatical resources influences linguistic complexity are in themselves quite complex.

Abbreviations

1	first person
3	third person
COM	comitative
COP	copula
EMPH	emphatic
F	feminine

FOC	focus
FUT	future tense
IMP	imperative
IPFV	imperfective
LOC	locative
NONFUT	non-future tense
OBJ	object marker
PL	plural
POSS	possessive
PST1	past tense 1
PST2	past tense 2
SG	singular

Text source

Nuevo Testamento. Mbia Cheẽ. (El Nuevo Testamento en Sirionó. Traducido por El Instituto Lingüístico de Verano. Riberalta, Bolivia. 1977).

References

Abbott, M. 1991. Macushi. In *Handbook of Amazonian Languages*, vol. 3, D.C. Derbyshire & G.K. Pullum (eds), 23–160. Berlin/New York: Mouton de Gruyter.

Dahl, Ö. 2004. *The Growth and Maintenance of Linguistic Complexity* [Studies in language companion series 71]. Amsterdam: John Benjamins.

Derbyshire, D.C. 1986. Comparative survey of morphology and syntax in Brazilian Arawakan. In D.C. Derbyshire & G.K. Pullum (eds), 469–566.

Derbyshire, D.C. & Pullum, G.K. 1986. Introduction. In D.C. Derbyshire & G.K. Pullum (eds), 1–28.

Derbyshire, D.C. & Pullum, G.K. (eds) 1986. *Handbook of Amazonian Languages*, Vol. 1. Berlin/New York: Mouton de Gruyter.

Dixon, R.M.W. 2004. *The Jarawara Language of Southern Amazonia*. Oxford: Oxford University Press.

Givón, T. 1979. From discourse to syntax: Grammar as a processing strategy. In *Discourse and Syntax*, vol. 12, T. Givón (ed.), 81–112. New York: Academic Press.

Kakamasu, J. 1986. Urubu-Kaapor. In D.C. Derbyshire & G.K. Pullum (eds), 326–403.

Koehn, E. & Koehn, S. 1986. Apalai. In D.C. Derbyshire & G.K. Pullum (eds), 33–127.

Stassen, L. 2000. AND-languages and WITH-languages. *Linguistic Typology* 4(1): 1–54.

PART II

Contact and change

Why does a language undress?

Strange cases in Indonesia*

John McWhorter
Manhattan Institute

I have argued in various presentations that it is inherent to natural grammars to maintain a considerable level of complexity over time: simplifications occur, but are counterbalanced by complexifications due to grammaticalization, reanalysis, and new patterns created by phonetic erosion. I argue that only extensive acquisition by adults makes grammars simplify to a significant overall degree. Creoles are the extreme case, but languages like English, Mandarin Chinese, Persian, and Indonesian are less complex than their sister languages to a degree that correlates with their extensive histories of non-native acquisition at certain points on their timelines. In this paper I address a few cases in Indonesia that challenge my stipulation. The grammatical simplicity of Riau Indonesian and the languages of East Timor is due to adult acquisition. Meanwhile, a few completely analytic languages on Flores suggest either that my stipulation must be taken as a tendency, or that we can take the nature of the languages as spurs for investigating sociological disruption in the past.

1. Introduction

I have argued (McWhorter 2001a) that creole languages are less grammatically complex than older languages. According to the taxonomy of complexity I present in the later McWhorter (2007), the differential consists of:

1. *overspecification*: marking of semantic categories left to context in many or most languages, such as evidential marking;
2. *structural elaboration*: number of rules mediating underlying forms and surface forms, such as morphophonemics; and
3. *irregularity*.

* I am eternally grateful to Wayan Arka, Mark Donohue, William Foley, and Geoffrey Hull for data, references and insights that have assisted me in venturing pronouncements about a language family in which I have no formal training. Special thanks are due to David Gil, with whom I have had the bracing pleasure of ongoing fruitful and civil exchange about Riau Indonesian and its relatives despite our differing conclusions.

I have also argued (McWhorter 2007) that this leads to a corollary thesis, that ordinarily, grammars maintain a high level of complexity over time, and if acquired natively by each generation, never drift into any overall simplicity. That is, it is impossible that a language would become as simplified as a creole simply by chance.

Certainly a grammar sheds complexities over time because of phonetic erosion and reanalysis. But at the same time, grammaticalizations and other reanalyses create new complexities. The inexorable nature of both processes is such that it is never the case that simplification predominates over complexification to any significant degree (cf. Dahl 2004: 11). To wit, there would seem to be no theoretical, or even logical, reason to suppose that it would: complexification is as natural in grammars' transformations through time as simplification.

The single factor known that would retard complexification is non-native acquisition. Pidginization and creolization are extreme manifestations of this. However, there are other cases showing that some grammars drift into relative, albeit not extreme, simplicity compared to their sister languages.

English is an example: as I have observed (McWhorter 2002), it is the only Indo-European language in Europe without grammatical gender, and the only Germanic language that lacks V2 word order, directional adverbs such as earlier English's *hither* and *thither*, a singular second-person pronominal, a verb devoted to indicating the passive such as German's *werden*, any productive use of the Proto-Germanic battery of derivational prefixes, and other features. While English has developed some new complexities of its own, these do not approach the volume of the losses – and we must also recall that the other Germanic languages have developed their own complexities as well.

Traditional analysis treats these abbreviations as just a pathway English happened to fall into, or at best, allows that the loss of grammatical gender was due to contact with Scandinavian Vikings, but treats the other features in mere descriptive fashion, apparently due to no unitary causal factor. In contrast, I argue that the entirety of these losses was due to extensive non-native acquisition by Scandinavian Vikings, such that the general contrast in complexity between Modern English and its sisters was determined by this encounter.

Similar arguments can be made, as I have presented in McWhorter (2007), for Mandarin Chinese compared to all of the other Chinese languages, Persian compared to Iranian languages, all of the regional Arabic varieties compared to Modern Standard Arabic, and Standard Indonesian compared to all of its closer Austronesian relatives. [1] All have been noted as peculiarly low in complexity compared to their sisters, and all of them have heavy non-native acquisition in their timelines, due to the vast population

1. Since writing McWhorter (2007) I have become aware that English was impacted significantly by Celtic languages, contrary to the traditional skepticism of this argument (McWhorter forthcoming). However, this *transfer* occurred amidst what remains an overwhelming amount of *simplification* unknown elsewhere in Germanic, and thus leaves my 2002 argument on English unaffected.

movements and empire-building that post-Neolithic agricultural technology conditioned, which underlay the emergence of massively populous civilizations, hungry for extra space and thus overrunning speakers of other languages.

I propose that heavy non-native acquisition is not merely one factor that can make a grammar drift into radical simplification, but that it is the sole factor. That is, I propose the following tenet for the relationship between grammatical complexity and language contact:

> In the uninterrupted transmission of a human language, radical loss of complexity throughout the grammar is neither normal, occasional, nor rare, but **impossible**. The natural state of human language is one saddled with accreted complexity unnecessary to communication. Wherever this complexity is radically abbreviated overall rather than in scattered, local fashion, this is not just sometimes, but **always** caused by a sociohistorical situation in which non-native acquisition of the language was widespread enough that grammar was transmitted to new generations in a significantly simplified form. This is true not only in the extreme case of creoles, but also to a lesser but robust extent in many languages of the world.

The idea that grammars can simplify vastly just by chance is conditioned in large part by the development from earlier Indo-European languages to modern ones, such as Old English to Modern English and Latin to Romance. However, in McWhorter (2007) I stress that these cases are in fact due to non-native acquisition themselves. These languages' central role in linguistic discussion is an epiphenomenon of the fact that academic inquiry is a product of civilizations borne of large-scale population movements, such that linguistic research has been spearheaded and dominated by speakers of languages with heritages in such sociohistorical developments, and languages of this type have tended to be recorded most in writing over time.

More illustrative of how conservative of complexity grammars are under normal conditions is a bird's-eye view of the world. There exists no mysteriously analytic Algonquian language. No Chinese language lacks tones. The only Semitic variety that eschews the family's hallmark triliteral consonantal template is, significantly, Nubi *Creole* Arabic.

I stipulate that any grammar that 1) has developed without a significant degree of non-native acquisition and 2) is starkly less grammatically complex than its sister languages constitutes counterevidence to my thesis.[2] In this paper, I will address a few cases that

2. Crucial corollary: languages that have come to be widely acquired as second languages only after standardization, and thus generally taught in school via the medium of writing, are not expected to simplify – or, their printed rendition will not indicate this. Prescriptive tendencies exert a conservative influence on the written language regardless of how the language is actually spoken casually. Thus, for example, Russian's widespread usage across the former Soviet Union has had no simplificatory effect on the written language and very well may never do so. The effect I refer to is one that occurs in languages before widespread literacy and standardization.

have come to my attention that could possibly be interpreted as such counterevidence. Two actually support my hypothesis, while one presents more of a challenge. For reasons unknown to me at present, all of these languages are spoken in Indonesia.

2. Why would a language undress?

Before addressing the Indonesian cases in question, it will be useful to be somewhat more specific as to my motivations for stipulating that grammars do not, under ordinary conditions, drift into vastly less complexity. The fact is that an unmediated, grammar-internal drift into analyticity manifests *no theoretically documented process of language change*.

Usually when a grammar changes in this way, linguists trace it to some kind of abrupt interruption in transmission due to population movements. For example, here is a sentence in an indigenous variety of the Bantu language Kikongo (1a) and then in the contact variety Kituba (1b), developed amidst speakers of various divergent Kikongo dialects amidst the upheavals of Belgian colonization. There is no Bantu variety in eastern or southern Africa remotely as isolating as Kituba – other than Fanakalo, a pidginized Zulu, or some varieties of Lingala, which is, again, a lingua franca that emerged amidst non-native acquisition (of the Ngala dialect complex):

(1) a. *(Yándi) ka-ku-zól-elé.*
he/she AGR-2S.OBJ-like-NEAR.PERFECT
b. *Yándi zóla ngé.*
he/she like 2s
'He/she likes you.' (Mufwene 1997: 176)

Scholars of language death associate such processes with interruption in transmission due to speakers failing to pass a grammar on fully to new generations.

Meanwhile, certainly grammars lose inflections to phonetic erosion. But they also develop new ones via grammaticalization and reinterpretation of phonetic segments. What factor, precisely, would condition a generation of speakers to reject a mass of, rather than just occasional, living features in the grammar they are exposed to as children *with no erosional sign left of the material shed*? If we suppose that the process occurred over several generations, then the question remains as to what would lead a first generation to reject even a good bundle of the living features their parents used? And then, if they for some reason did, why would the succeeding generation take this as a "trend" and continue this rejectional orientation towards their native grammars, to a degree significant or even slight? What theoretically constrained factor would prevent this succeeding generation from simply reverting back to "normal" and reproducing the grammar as they learned it?

For example, if English evolved into an entirely analytic language, all theories of language change would identify non-native learning as the reason – and indeed, the

only descendants of English with, for example, no inflection at all are creole languages in Surinam like Sranan and Saramaccan.

3. Riau Indonesian

Gil (2001), in response to my argument about creoles' lesser complexity than older languages, presents the nonstandard Indonesian dialect of Riau province in Sumatra as an exception. Gil certainly shows Riau Indonesian to be an almost counterintuitively telegraphic form of natural language. Its word order is free. There are no copulas, complementizers, or relativizers. There is no possessive marker according to Gil's analysis. While in Standard Indonesian, use of tense and aspect markers is optional, in Riau they are even less conventionalized. Numeral classifiers are an areal feature and used in Standard Indonesian, but rarely in the Riau variety (ibid. 343–357).

Riau Indonesian also recasts Standard Indonesian's voice-marking morphology – agent-oriented *meN-* and object-oriented *di-* – into optional pragmatic markers of agent and patient (Gil 2002). Thus whereas in Standard Indonesian, *di-* marking connotes a passive reading, in Riau Indonesian it only lightly highlights patienthood:

(2) *Aku* ***di****-goreng.*
I oo-fry
Standard: 'I was fried.'
Riau: 'I fried it.' (or by my reading of Gil, 'I fried the thing.') (Gil 2002)

The Riau variety is but one of several colloquial Indonesians that Gil calls attention to in which the situation is similar, such as one in southern Sulawesi, Kuala Lumpur, and Irian Jaya (which makes barely any use of either voice marker). Overall, Standard Indonesian is the "high" member in a diglossia, an artificially constructed language few speak casually. Gil describes Indonesian as it has actually evolved as a spoken language.

Gil (2005) states that "the accretion of complexity cannot be construed as an inexorable monotonic process," since apparently "at some stage between Proto-Austronesian and Riau Indonesian, the accretion of complexity must have been reversed." But when he notes that this leaves a question as to "why more languages could not have taken the same path", and leaves it at that "I have no answer to this question", I propose that the reason is that no languages have taken this path – that is, without extensive non-native use.

I propose the following arguments for this position.

3.1 There are no varieties of Indonesian's relatives that are reduced like Riau Indonesian

If the nature of Riau Indonesian plus several other colloquial Indonesian varieties were a typical result of uninterrupted transmission, then it would be nothing less than common to find similarly reduced varieties of languages closely related to Indonesian like

Javanese, Sundanese, Madurese, Minangkabau and Iban. Given that these languages have similar grammars to Indonesian's, then surely happenstance would exert like effect upon at least a few varieties of some of these languages.

This is, to my knowledge at this writing, unknown. A relevant example is Smith-Hefner's (1988) description of the Javanese of Tengger, differing from the standard in minor elisions and collapsings of morphology, but hardly to the extent of Riau Indonesian from the standard.

Importantly, we would not need to assume that all such cases would be, like the spoken Indonesians, socially unacknowledged colloquial levels of more complex standard languages, difficult to smoke out without fluent command of the language and thus unsurprisingly undocumented today. Since the vast majority of Indonesian languages are small, unwritten ones, if Riau Indonesian's nature is a natural development, then we would expect that a map of Sumatra or Sulawesi would be dotted with at least occasional highly analytic languages that had happened along the same pathway, with no artificially preserved written variety to mask the spoken reality.

At our current state of knowledge, there are no such varieties. I submit that this is because under conditions of full transmission, natural languages *never* evolve in such a direction.

In an ingenious study, Gil (to appear) proposes that Indonesian's close relatives Minangkabau and Sundanese can be shown to have indeed reached a radical degree of reduction despite not having histories in which they were learned more by adults than children. Gil's metric is the freedom of interpretation of items under association. In a language like English, association is highly constrained by rules such as subjects normally preceding objects, such that if an English-speaking subject is presented with the sentence *The bird is eating* and then shown one picture showing a bird eating and the other showing a cat eating a bird, he or she will point to the former picture. However, if a Sundanese speaker is given the equivalent sentence *Manuk dahar* 'bird eat,' then they will accept both pictures as illustrating the sentence, because it could also mean 'He (the cat) eats the bird.' The absence of case or inflectional affixation constraining the interpretation makes this possible in Sundanese.

Testing sixteen sentences, Gil shows that of eleven languages, Minangkabau and Sundanese have the freest association according to this metric – and, by implication, parallel Riau Indonesian, whose freedom of associational semantics Gil has examined elsewhere.

It is certainly true that compared to the hundreds of other languages of the Western Malayo-Polynesian subdivision of Austronesian, Minangkabau and Sundanese stand out as relatively low in grammatical complexity, lacking, for example, tense and aspect affixes or concordial affixes on the head. However, internal evolution is not the only possible cause here. In this, Minangkabau and Sundanese parallel Standard Indonesian, and in McWhorter (2007) I suggest that this is an areal trait due to the intimate contact with Indonesian that these languages (as well as Javanese and Madurese) have had over the eons. The associational freedom that Gil analyses can be seen as due to this reality.

Moreover, the fact remains that this associational freedom is but one aspect of what grammatical complexity can consist of, and in many other areas of grammar, Minangkabau and Sundanese surpass Riau Indonesian. For example, they parallel Indonesian's battery of derivational morphology complete with morphophonemic complications in their surface realization (something the sentences in Gil [2007]) do not happen to cover). Meanwhile, Minangkabau surpasses even Standard Indonesian in grading distance more finely and in the number of numeral classifiers (Moussay 1981). Sundanese has the same traits, as well as productive infixation, a negative existential marker, a high yogh vowel along with the typical vowel inventory of languages of its Malayo-Javanic group, and other features (Müller-Gotama 2001).

Thus I accept Gil's argument that his study shows that a language can drift into extreme simplicity in a particular area of its grammar – as well as that in this case, Minangkabau and Sundanese clearly surpass even creoles like Saramaccan. However, in that my interest is in overall complexity, it would seem that the study does not, in itself, demonstrate that Minangkabau and Sundanese have by chance drifted into the *overall* degree of simplicity of Riau Indonesian.

3.2 The grammar of Riau Indonesian gives no concrete indication of transfer from other languages

Gil also argues that Riau Indonesian's nature is due to a Sprachbund feature typical of Southeast Asia, a tendency to allow considerable ellipsis of arguments. However, the fact remains that languages with this tendency can be inflected nevertheless, such as the Philippines languages and so many other Austronesian languages, or even if analytic, have ergativity (like Polynesian languages) or massive lists of numeral classifiers (common in Southeast Asia in general). Our question, then, is why Riau is so naked of complexity even beyond its tendency to elide arguments unnecessary to comprehension?

The question is why Riau Indonesian apparently incorporated only this abbreviational feature of surrounding languages, rather than concrete additions from them. A contrasting case is Acehnese, the result of a language closely related to Indonesian mixing with Mon-Khmer:

(3) *Na si-droe-ureueng-ladang geu-jak lam-uteuen geu-jak koh kayee.*
be one-CLAS-person-farm 3s-go in-forest 3s-go cut wood
'There was a farmer who went into the forest to cut wood.' (Durie 1985: 192)

The mixture here is unequivocal. The phonetic inventory is unlike that in any related language. Orthographically, *eu* represents unrounded back vowel [ɯ]; *eue* represents a diphthong with a schwa offglide ([ɯə]); *oe* represents [ɔə]. These are clear parallels to typical sounds in Mon-Khmer languages. The *droe* classifier comes from the use of this word as 'self' (cf. Indonesian *diri*), but Indonesian's cognate is not used in this function. The syntax departs from Indonesian as well, with its obligatory subject-marking clitics and absence of voice-marking prefixes (*di-* and *meN-* have no cognates or equivalents in Acehnese).

Importantly, however, there is nothing especially "simplified" about Acehnese overall, a well-inflected and overspecified grammar. Here, then, is a relative of Indonesian's in which mixture has created a language sharply different from any form of Indonesian itself. Riau Indonesian, obviously, is not mixed to anything approaching this extent. It differs from Standard Indonesian in its radical simplification.

3.3 Theories of grammar-internal change do not predict the fate of voice markers in colloquial Indonesians

The drifting of semantic markers into more optional and less regularized pragmatic ones goes against the strong tendency that grammaticalization specialists recognize in the other direction. This renders the fate of *di-* and *N-* in colloquial Indonesians peculiar – until we recall that non-native acquisition can lead to precisely this kind of reversal.

In creoles, for example, categories obligatorily and even redundantly expressed in source languages are very often marked in a smaller range of contexts in the creole, and often only for explicitness. Palenquero Creole Spanish inherits the *ma-* plural affix from Kikongo. But whereas Kikongo marks the plural categorically as does Spanish, Palenquero, besides "loosening" the affix into the less bound status of a clitic, often omits it when context supplies the inference:

(4) a. *Kasi to* ***ma*** *moná di ayá la baho*
almost all P girl of there there below
'Almost all the girls from down there'

b. *Esa ea* ***ø*** *mamá puñera-ba re akí.*
that be.PAST mother boxing-PAST of here
'Those were the boxing women here.' (Schwegler 1998: 289)

In cases like pidginization and creolization, "unravelling" of this kind is ordinary. In uninterrupted language change, it is anomalous.

3.4 Summary

On the basis of the above observations, I suggest that Riau Indonesian is not a natural development that Indonesian grammar would take, and that explaining this grammar as a Sprachbund phenomenon is only partially useful in the explanatory sense. Riau Indonesian gives all evidence of being what happens when Indonesian is acquired widely enough non-natively that the complexities of its grammatical machinery start falling away in cases where context can easily convey their meaning or function.

It is important that as Gil notes (2001: 330–334), even today one in four Riau Indonesian speakers grew up in homes where at least one parent was not a native Indonesian speaker, and that "the present-day Riau province was the venue of substantial language contact over much of the last 2000 years", that "various contact varieties of Malayic must have arisen during this lengthy period", and that "such contact varieties

constitute plausible ancestors for what is now Riau Indonesian." In general, only 7% of Indonesians speak Indonesian as a first language (Prentice 1990: 915).

To be sure, this proposal is only that; there exists no sociohistorical data remotely detailed enough to decisively conclude that it is accurate. Gil (p.c.) notes that his one-in-four figure is based on his informal investigation, and that sociohistorical documentation on the development of Riau Indonesian is likely unavailable. However, what Gil discovered in his survey can be plausibly taken as germane to the uniquely streamlined nature of the grammar. This is because, precisely, there is no known diachronic mechanism that would yield such results other than extensive second-language acquisition.

4. Timor

At first glance, Timor would appear to be a handy demonstration of my thesis, in the difference between the two main varieties of Tetun. Tetun Terik is spoken only in a southern coastal region, while Tetun Dili is a lingua franca of the island, spoken by two-thirds of the population, whose native languages are eleven other Central Malayo-Polynesian languages and four distantly related to Papuan ones.

4.1 Tetun Terik vs. Tetun Dili

Tetun Dili (TD), predictably according to my thesis, has a saliently simpler grammar than Tetun Terik (TT). (Following data from Van Klinken 1999 and Williams-Van Klinken, Hajek and Nordlinger 2002.) TT has three verbal affixes: *ba-* for causative, *hak-* as an intransitivizer, and *mak-* for marking actorhood (*daka* 'guard', *ema mak-daka-n* 'person who guards'). In TD, *mak-* is absent, and the *hak-* cognate is present but less obligatory or productive than in TT. TT uses six numeral classifiers. In TD nouns usually occur without classifiers, and only four are in use at all. TT has a definite article, although its occurrence varies geographically; TD leaves definiteness to context. In TT, possession entails enclitics on the noun whose choice is determined by constraints of alienability and other factors: *ó fé-**r*** or *ó fé-**n*** 'your wife,' but *sa ata-**r***, **sa ata-**n*** 'their slave.' TD has lost the elaborate conditioning of these clitics, instead using a free morpheme *nia* for possession (*ahau nia liman* 'my hand'), with fossilized *-n* as the only remnant of the TT situation (*uma João nian* 'John's house'). TT has three functionally distinct copulas; TD retains just one (*iha*). TT has an equative negator and a predicate one:

(5) a. *Buat e'e Bei Beur **ha'i!***
thing this Mr. deceive NEG
'This thing isn't Mr. Trickster!'

b. ***la** bele la'o*
NEG can walk
'cannot walk'

TD retains just *la* across the board. Most famously, TT has subject-marking concord prefixes. Their appearance is complexified by morphophonemic rules. For example, the prefixes replace initial /h/, such that ***k**-sai* 'I exit' but with *há* 'to eat', *ká* 'I eat' rather than **k-há*. When the *h*-initial verb is trisyllabic and begins with the *hak-* prefix, the prefix's /k/ is optionally elided in the first person singular: *hakdiuk* 'play' > *k-akdiuk* or *k-adiuk*. TD, however, lacks subject-marking prefixes entirely, and is largely an isolating language.

4.2 Tetun Terik vs. its relatives

However, in wider view, Timor gives us more to chew on than this. Even Tetun Terik is strangely streamlined compared to typical members of its Central Malayo-Polynesian group.

For example, Rotinese (Timor and Roti) and Kambera (Sumba) have subject-marking prefixes like Tetun Terik, but otherwise surpass it in general complexity. Rotinese has eight numeral classifiers and encodes the possessive with enclitics conditioned by constituent class (Jonker 1915: 270–274). Kambera marks subjects with the accusative in impersonal constructions, thus rendering an absolutive distinction in a language otherwise nominative/accusative:

(6) *Jàka nda nyumu, meti-**ya**.*
CONJ NEG you die-3S.ACC
'Without you, we would die.' (one would have died) (Klamer 1998: 161)

(7) *Nggàra mài-**ya**-i nú?*
what come-3S.ACC-ITER there
'What can I do for you?' (ibid. 167)

It is important to note that subject-marking prefixes can complexify not just in overspecifying the subject, but in being distributed into semantically arbitrary conjugational classes and occasioning morphophonemic irregularity. Here are verb paradigms in Sika of Flores Island, in which prefixes have eroded into consonant mutations:

Table 1. Verb paradigms in Sika (Lewis and Grimes 1995: 605–606)

	***bano* 'go'**	***gita* 'see'**	***raʔit* 'sew'**
1S	*pano*	*ʔita*	*ʔraʔit*
2S	*bano*	*gita*	*raʔit*
3S	*bano*	*gita*	*raʔit*
1P (exclusive)	*bano*	*gita*	*raʔit*
1P (inclusive)	*pano*	*ʔita*	*raʔit*
2P	*bano*	*gita*	*raʔit*
3P	*pano*	*gita*	*raʔit*

Verbs in Dawanese change shape in several fashions (including metathesis) according to factors such as presence and placement of object and whether a following object begins with a consonant cluster (ibid. 141–142). Thus with no object:

(8) *Qau* ***qòt***
I cut
'I cut.'

With a fronted object:

(9) *Nuif qi qau* ***qote***
bone this I cut
'This bone I cut.' (ibid. 142)

A following object triggers metathesis:

(10) *Hò m-**qoet** nuif?*
you 2s-cut bone
'Do you cut bones?' (ibid. 141)

Grammars like Sika and Dawanese, under my analysis, are business as usual. On the Baikenu dialect of Dawanese, for example, Hull (2001a) in a popular source has it that "unlike Tetum, which is a structurally simple language and easy to learn, Baikenu is relatively difficult," such that "for it to become more accessible to foreigners, it urgently needs more comprehensive pedagogical resources than a simple language manual" – i.e., Dawanese is like most languages, while something has made Tetun more user-friendly. This leaves a question as to why even the indigenous variety of Tetun, Tetun Terik, has so much less complexity.

One might propose that Tetun Terik has somehow undergone "blowback" effects from the reduction in Tetun Dili: many speakers may have been competent in both varieties and then, since they are so similar, began mixing the grammars, such that Tetun Terik became intermediately complex between a language like Dawanese and a variety like Tetun Dili.

But this runs up against another problem. Under that analysis we would suppose that all of the other indigenous languages of Timor were riddled with complexity, but in fact, a certain anomalous simplification is found in all of East Timor's languages:

Table 2. 'They stole my buffalo and killed it' in four Timorese languages (from Hull 2001b: 163)

Tetun Terik	*Sira*	*na'ok*	*ha'u*	*karau,*		*roho*	*tiha.*
Galoli	*Sia*	*na'o*	*ga'u-ni*	*karau*	*no*	*regen*	*oin.*
Tokodede	*Ro'o*	*mana'o*	*aka'u*	*karbau,*		*hora*	*bali.*
Mambai (Ainaro)	*Rom*	*pnao*	*au*	*arbau,*		*dae-pul*	*tel.*
	they	steal	my	buffalo	and	kill	PERF

Tokodede, in particular, is as analytic as Tetun Dili (Hull 1998: 166, 2001b: 101).

The import of there being a complex of grammars of this kind on one island is how atypical this homology is of Austronesian languages. Philippines-type languages require focus marking on one constituent via the combined application of affixation and a free morpheme. Most other Western Malayo-Polynesian languages (such as Indonesian, and thus often referred to as "Indonesian-type" languages since more detailed taxonomy currently remains to be done) often have concordial marking for subjects and/or objects, as well as much more elaborated marking of grammatical relations, tense, negation, etc., or, like Indonesian, grammaticalized voice marking (cf. Donohue 1999 on Tukang Besi). Most remaining Austronesian languages are in the Oceanic group, which is typified not only by healthy amounts of inflection, but two differently marked classes of possession, robust marking of transitivity, a proliferation of number and inclusivity distinctions marked by pronominals, arguments coreferenced with affixes on the head, etc., or in Polynesian, marking of grammatical relations with particles, including sometimes ergativity.

Then we return to Central Malayo-Polynesian languages of normal complexity like Kambera, Rotinese, and Dawanese. What traditionally known process would mysteriously lead some grammars to devolve from typical Austronesian complexity to Tokodede's telegraphic nature? Why does a grammar replace not just some, but almost all bound constructions with periphrastic ones?

4.3 Things just happen? The Papuan factor

Of course, at this point one might conjecture that the data are telling us that for some reason, an ancestral language at some intermediate point in Central Malayo-Polynesian simply opted to "take it all off" by chance, and passed on this trait to its descendants on the eastern half of Timor. But as it happens, there are four languages spoken in East Timor that are Papuan in origin (Timor being the westernmost boundary of the Papuan languages' area). Bunak, Makasai, Makalero and Fataluku show the same unusual amount of simplification as East Timor's Austronesian languages. They are, of all things, analytic Papuan languages, which are otherwise notoriously morphologically elaborated (Foley 1986: 12), typically of ancient languages spoken by small groups, in which if there is not ample inflection then there tends to be equivalent tonological elaboration, as in Southeast Asian and West African languages. This clearly refutes a mere genetic explanation, in that these languages are not even of Austronesian descent.

Since the Papuan group includes about 800 languages, certainly some of them happen to be analytic. However, this is extremely rare: in Foley's entire survey of the group, for only one language, Asmat, are there citations that happen not to display inflectional morphology, and this is due merely to chance, as Asmat is highly inflectional (cf. Drabbe 1963). Foley (p.c., March 2006) is aware of only a few isolating Papuan languages, in the northeastern region. The Papuan languages of Timor, on the other hand, have their source languages on Irian Jaya's Bomberai peninsula in the northwest (cf. Hull 1998: 21–23 for a summary of research on the issue), and these languages,

such as Iha, Baham, and Mor, are highly inflected (Mark Donohue, p.c., March 2006). Thus there is no evidence of an analytic group of Papuan languages, or single source language, imported to Timor. Rather, inflected languages were brought to Timor and then mysteriously drifted into analyticity. In contrast to typically inflectional Papuan languages is a Fataluku sentence such as *Ana merkadu mara* (I market go) 'I'm going to the market' (Hajek and Tilman 2001: 177). Behold a Papuan language with the typology of Chinese.

It would be unconstrained to suppose that this were a mere matter of influence from Tetun Dili or the Timorese languages generally. We must explain why these four languages have not just lost a certain amount of inflections, but why they underwent a decisive typological transformation such that their SOV word order stands as one of the only features they retain in common with their Papuan ancestors, almost all of which are highly inflectional.

An account based on mere long-term bilingualism with an isolating language (or languages) becomes especially incoherent in view of the fact that so very many languages of Papua New Guinea and islands eastward have been spoken alongside analytic languages like Tok Pisin and its sister creoles for a century and a half. Guinea-Bissau Creole Portuguese has been spoken for five centuries by speakers of West Atlantic languages, but their renditions of Balanta, Manjaku and other languages remain as complex as they have always been. Agglutinative Altaic languages have been spoken alongside Mandarin Chinese for millennia, and yet the result is not analytic Altaic languages, but highly agglutinative hybrid dialects (cf. Lee-Smith 1996; Lee-Smith and Wurm 1996) of Mandarin and Altaic.

Why not, then, at least one or two of the four Papuan languages of Timor in which Papuan morphology remains, but is remodelled partially upon the morphological categories of conservative varieties like Tetun Terik? Why not Papuan languages where most of the morphology remains, but is only slightly eroded by contact?

4.4 Just chance?

Some might suppose that it is scientific to simply assume that the languages of East Timor just happened to drift into analyticity the way others drift into ergativity. But this returns us to the fact that a grammar-internal drift into analyticity manifests no theoretically documented process of language change. Over countless millennia, not a single Indo-European language has lost so much affixation as to become typologically akin to Chinese, nor is there a fortuitously analytic Uralic, Caucasian, Dravidian, Paleosiberian, or Algonquian language.

In West Timor or East Flores, at the front and rear end of verbs, erosion and effects of millennia of contact with pronominals have left complications as shown above in Dawanese and Sika: this is "normal." In East Timor, for some reason affixes just vanished and left roots unaffected – i.e., without conjugational alterations as reflections of bygone affixations and abutments. This is not "normal" – it suggests that learners

were acquiring the system incompletely, abruptly tossing away machinery in a fashion alien to how uninterrupted language change works.

Moreover, the loss of complexity in East Timor's languages goes beyond inflection. Tukang Besi has at least twelve numeral classifiers (Donohue 1999: 109); Tetun Terik has just six. Tukang Besi has distinct pronominal paradigms for tonicity, realis, irrealis, possessive, object and dative (ibid. 113), while Tetun Terik knows nothing of the sort; and so on. Just as it is difficult to conceive of why speakers of a language would toss out almost all of its inflections without replacing them, no theoretical schema can explain why a generation of speakers would, for example, stop using most of a language's numeral classifiers or eliminate case distinctions in pronominals.

Thus in view of my thesis that ordinarily, languages do not simply shed most or all of their complexity and develop grammars as simplified as ones born a few hundred years ago from pidgins, we must hypothesize that something unusual happened to the languages of Timor – probably having to do with contact. Something did.

4.5 What happened in Timor?

There is a unique and strong lexical and grammatical imprint in these languages from Central Malayo-Polynesian languages spoken north of Timor, on the Uliasser islands – Ambon and adjacent islands – and to a lesser extent, on nearby Ceram, Buru, and Ambelau. The languages include Asiluluan, Hitunese, Larikean, Batumeran, Harukuan, Saparuan, and Bahasa Nusa Laut of the Uliassers, Nialan, Bonfian, Warunese of Ceram, and Masarete and Kayeli of Buru (Hull 1998: 154). Hull refers to this source generically as an "Ambonic" imprint.

The lexical legacy in Timor includes about a hundred items, of core semantics rather than "cultural" borrowings, including grammatical items such as 'no,' prepositions such as 'at' and 'to', and adverbs such as 'before' and 'tomorrow' (Hull 2001b). Then the languages of Timor also have various grammatical features which together point to these Ambonic languages – rather than, for example, the Sulawesi languages such as Tukang Besi and Muna from which Hull reconstructs the Timoric languages as having developed – such as an inalienability distinction, an agentive prefix, a negative imperative marker, postposition of numerals and preposing of possessive markers and a great deal more (Hull 1998: 166, Hull 2001b: 115).[3]

Given the bizarrely analytic typology of the languages of East Timor, this applying to languages of not one but two groups, it is plausible to assume that this Ambonic impact occasioned extensive and abrupt non-native acquisition of these languages. The extent

3. Hull hypothesizes that the languages of Timor are descendants of the Western Malayo-Polynesian languages mentioned, and contests the Central Malayo-Polynesian classification. I will sidestep this issue here, as it is not germane to the presentation and goes beyond my qualification to judge.

and depth of the lexical borrowings, for example, are analogous to the Scandinavian ones in English, which were part of an encounter that led English to shed more of the Proto-Germanic legacy than any language in its subfamily (and as I have argued in McWhorter 2002, there are Scandinavian grammatical imprints on English as well).

As such, Hull (1998) reconstructs that in approximately the 1200s AD, invaders from the Ambonic-speaking region disrupted language transmission where they settled. Evidence beyond the linguistic includes that Timorese origin myths correspond to lore connected with the Moluccan aristocracy (Ambon and surrounding islands are part of the Molucca group) rather than to that connected with Sulawesi noblity; that Rotinese tales of origin describe inmigration by a band of Ceramese over 400 years before the arrival of the Portuguese; and that speakers of the Timoric language Galoli recount that in antiquity, men from Nusa Laut arrived and married local women (ibid. 161–164). Especially indicative of intimate interactions between Timor and Ambon is that six place-names on Ambon correspond to Timorese groups (Hull 1998: 162).

While the sociohistorical details of the Ambonic migration are unrecoverable, I suggest that treating this migration as the cause of the strangely low level of complexity in Timor languages' grammars is more scientific than ascribing the anomaly to chance. This is for the simple reason that theories of language change offer us no mechanism or motivation for a language to shed its complexities to any significant degree amidst full transmission to successive generations.

5. Flores

The oddest cases of decomplexification known to me are a few Austronesian languages of Flores: specifically, Keo, Rongga, and Ngadha. These languages are completely analytic, without even the subject-marking prefixes or small collection of derivational prefixes typical in Timorese languages. Here is a Keo sentence (Baird 2002: 491):

(11) *Ho ghonggé ndé, dima ʻimu mélé ka ʻuru néʼé*
oh grope.in.hole that arm 3s stuck PERF because have
kara subho
bracelet shell bracelet
'When she was groping in the hole, her arm got stuck because she had a shell bracelet.'

(I assume that at least some of the virtually unexamined languages in the surrounding area are likely of similar typology.)

Austronesianists treat these languages' analyticity as a matter of chance, of little inherent interest. But under my analysis, these languages are utterly bizarre. Why did a descendant of Proto-Austronesian wend its way into eschewing all affixation?

5.1 Complexity without affixation

As it happens, the challenge that the Flores languages present to my thesis is very specific: their analyticity compared to other Austronesian languages. Affixation is, in grand view, but one of a great many manifestations of grammatical complexity, and these Flores languages are hardly devoid of these latter. Before proceeding, this must be made clear, in order to avoid the misimpression that these languages contradict my claims elsewhere that creole languages exhibit an unusually low degree of complexity (McWhorter 2001a).

Keo, for example, has 24 sortal classifiers, such as *dudhu* for mixtures of small things like cut-up vegetables, '*éko* (tail) for land-dwelling animals, '*étu* (meat) for kinds of meat, '*isi* for root vegetables, *mboko* for round things, *pata* for flat, thin objects, '*ula* for long, thin objects, and so on (Baird 2002: 253):

Table 3. Some Keo classifiers (Baird 2002: 243)

*muku '**esa** rua*	'two bananas'
*muku **sepi** rua*	'two bunches of bananas'
*muku **guni** rua*	'two stalks of bananas'
*muku **wunu** rua*	'two banana leaves'
*muku **pu'u** rua*	'two banana plants'
*muku **doka** rua*	'two clumps of banana plants'

Numeral classifiers, in subdividing nominals into classes, can be taken as a different manifestation of the role that grammatical gender marking plays in other grammars (Grinevald and Seifart 2004). Grammatical gender tends to additionally entail concord between heads and modifiers, but this is ultimately one end of a cline of development.

Keo, then, has "gender" (cf. Kihm 2003 on a definition of inflection independent of boundness), as do Rongga (Arka 2006 lists 22 sortal classifiers) and Ngadha (Arndt 1933: 6–7 lists 19). In Ngadha, just as grammatical gender marking tends to drift into semantic arbitrariness, some classifiers apply to unpredictable assemblages of objects: *vid'a* is used with books, trees, and pieces of meat; *mata* is used with fields, (water) springs, and tree trunks (ibid. 6).

Another complexity (overspecification) in Keo is that possessive marking differs according to alienability:

(12) a. *'udu nga'o lo*
head 1s hurt
'My head hurts.' (Baird 2002: 210)

b. *'aé **ko'o** kami*
water POSS 1P.EX
'our water' (ibid. 204)

Yet the inalienable particle is used under certain conditions with kin terms:

(13) *'ana* ***ko'o*** *weta*
child POSS sister
'sister's child' (ibid. 214)

Possession also differs according to alienability in Ngadha (Arndt 1933: 8).

While analytic, Ngadha has a proclitic that marks objects when pragmatic emphasis upon it is indicated, in a fashion reminiscent of the "object-oriented" affixes in Indonesian and its relatives (Arndt 1933: 9):

(14) *maë vela cata.*
NEG kill person
'You shouldn't kill anyone.'

(15) *gazi vela* ***go*** *radza.*
3S kill OBJ king
'He murdered the king.'

Meanwhile, Ngadha makes ample use of modal particles to convey pragmatic meanings, in a fashion similar to German's. Certainly all natural languages can convey such pragmatic meanings, but where grammars develop segmental indications of such beyond ordinary mechanisms of intonation or phrasal collocations, it qualifies as a grammatical overspecification. For example, here is *dzaö laä* 'I go' modified with just a few of the particles:

Table 4. Some Nghada modal particles (Arndt 1933: 27; trans. from German mine)

dzaö ***mu*** *laä*	'I'm going, whatever happens.'
dzaö ***mu le*** *laä*	'I'm keeping on walking along, whatever happens.'
dzaö ***mu mara*** *laä*	'I'm going whatever happens, and I'll deal with the circumstances.'

Overall, the tendency to associate complexity with morphology overlooks the considerable degree of complexity that free morphemes can lend. For example, here is a sentence in the Sino-Tibetan language Akha, in which there are markers of ergativity, an evidential marker, and a quotative marker in a thoroughly analytic language:

(16) *Àsjhaŋ sjháŋ nɛ shḿ-thö nɛ bzö sèq ŋà djé nà-bɔ́ áŋ nɛ.*
A. S. ERG awl INST hole kill EVID QUOT ear LOC INST
'Asjhang Sjhang killed her with an awl through her ear.' (Hansson 1976: 16)

And then, suppletivity can lend complexity: Ngadha, for example, has some cross-linguistically unusual suppletive negated verbs, such as *nge* 'to be able' and *talo* 'to not be able;' *beö* 'to know' and *busa* 'to not know' (Arndt 1933: 10).

Thus according to the conception of complexity that I present in McWhorter (2001a), Keo, Rongga and Ngadha are readily distinguishable from creoles in terms of overall complexity. Certainly, a given complexity that they have may not be completely unknown in any creole: e.g., Saramaccan has an incipient (albeit highly variable)

marking distinction between alienable and inalienable possession. However, in terms of degree, the Flores languages are more complexified than any creole known to me.

Nevertheless, the fact remains that these languages of Flores shed a significant battery of complexity in eschewing paradigms of affixes still living in the 1000 other languages of their family. Affixation leads to complexity in particular with its tendency to condition morphophonemic alterations and irregularities, of the sort shown previously in this paper in Indonesian, Sika, and other languages. Why did these Floresian descendants of Proto-Austronesian drift into the homology of a language like Akha? It is one thing to suppose that "a grammar might become analytic," and another to propose what the intermediate steps would be between, for example, Tukang Besi and Keo, and what would motivate them.

5.2 A phonological explanation?

I am aware of but one systematic account of how a grammar might transition from syntheticity to analyticity independently of language contact, Hyman's (forthcoming) hypothesis regarding the analyticity of West African branches of Niger-Congo in contrast to the prolifically agglutinative typology of most Bantu languages. A natural assumption would be that if some Niger-Congo languages like Yoruba are highly analytic:

(17) *Mo mú ìwé wá fún ẹ.*
I take book come give you
'I brought you a book.'

then this was the original state of Proto-Niger-Congo, and that typical Bantu verbal constructions like East African Yao's:

(18) *taam-uk-ul-igw-aasy-an-il-a*
sit-IMP-REV-PASS-CAUS-REC-APP-V
'to cause each other to be unseated for'

must be a result of free words falling into morphologization.

But Hyman proposes that Proto-Niger-Congo was an agglutinative language along the lines of today's "classic" Bantu, and that languages like Yoruba developed when agglutinative ancestors took on phonological templates that restricted verbs' syllable count, such that only bi- or trisyllabic words were allowed. As such, Niger-Congo languages in Cameroon, West Africa, and along the northern boundary of the Niger-Congo realm are, if not as strictly monosyllabic as Yoruba, limited to less syllabically prolific phonological templates. Thus take Yao's capacity:

Table 5. Morphology of the Yao verb

taam-a	'sit'	
taam-ik-a	'seat'	impositive
taam-ik-ul-a	'unseat'	reversive
taam-ik-ul-igw-a	'to be unseated'	passive
taam-ik-ul-igw-aasy-a	'cause to be unseated'	causative

An Edoid language of Nigeria like Degema has no verb monsters like Yao, but allows a causative marker cognate with Yao's *-assy-*, as in *tul* 'to reach', *tul-ese* 'to cause to reach.' The Ubangian language Banda-Linda of the Central African Republic has a reversive marker cognate with Yao's (with an alternation of Yao's lateral consonant to a rhotic): *ʒe* 'bubble, overflow, belch,' *ʒè-rə̀* 'deflate, take a final breath, shove into, go down.'

Along these lines, we could consider whether Keo is an "unravelled" version of an inflectional ancestor, having settled upon, for example, a disyllabic template.

But Hyman's analysis leaves a question: if it is so natural for a language to develop a phonological template so incommensurate with its typology that its entire grammatical structure is transformed, then why has no Bantu language in southern or eastern Africa drifted into such a phonology over the 3000 years since Bantu speakers started migrating from Cameroon? For that matter, we also seek similar examples in agglutinative languages around the world, i.e., the Turkic or Eskimo-Aleut language that limits affixation to one or two morphemes per root.

And again, it is difficult to imagine why a certain generation would begin contracting the number of syllables a word was allowed to have: the process is easier to describe than to actually explain. It is more graceful to suppose that Niger-Congo was indeed more like Yoruba and that Bantu languages are a particular branch in which an extreme degree of agglutinativity has developed. This, after all, would have happened via familiar processes of grammaticalization – i.e., it is clear and documented how this kind of change would occur across generations. Crucially, it is equally understandable why once this agglutinativity had emerged, it would persist in hundreds of languages over millennia.

We imagine that the next natural step would be not phonological undoing of boundness, but fusional typology; i.e., boundness increasing to the point of phonetic collapses. For example, some agglutinative morphemes would erode – but not into the ether, but leaving behind vocalic alternations upon other morphemes, yielding constructions like the ablaut patterns in families like Indo-European, much older than Bantu. And then a further, or alternate, step would be the development of tonal morphology as in Chinese.

Thus in my opinion, we cannot attribute Keo, Rongga and Ngadha's typology to templatic transformations in phonology.

5.3 A Wave of Simplification?

In a referee report, Wayan Arka has pointed out that there is a cline of increasing analyticity across the island chain from Bali to Flores, which could be interpreted as indicating that simplification indeed occurred naturally. The idea here would be that one language happened to shed a certain degree of inflection in becoming another one, and that then its descendant shed some more, and so on, until a language like Keo was the end of the chain. As such, Keo would represent not a single massive shucking of inflectional affixation, but just the end of a stepwise progression that was never abrupt in any sense.

Arka gives examples from Bima, spoken on Sumbawa just westward of Flores. In Bima, there are tense-marked sentences with person-tense portmanteau bound forms, such as:

(19) *Nahu* ***ka****-hade* *sawa* *kamau* *ake*
1s 1s.FUT-kill snake rice.field that
'I will kill this python.'

as well as tense-specific passive bound morphemes:

(20) *Oha* *ede* ***di****-ngaha* *ba* *ama* *nahu.*
rice that PASS.FUT-eat by father 1s
'The rice will be eaten by my father.'

However, there are also analytic sentences such as this one in which past semantics are rendered by context:

(21) *Sia* *hade* *sawa* *kamau* *ede.*
3s kill snake rice.field that
'She killed the python.'

and the passive can also be conveyed with a bare verb:

(22) *Sawa* *kamau* *ede* *hade* *ba* *sia.*
snake rice.field that kill by 3s
'The python was killed by him/her.'

However, for one, bare verbs are also common in Standard Indonesian, in which tense and aspect marking is often left to context. Thus this particular data does not indicate a drift towards analyticity from west to east.

And the question regarding the Flores languages is the one I pose in the title of this paper: why – or more to the point, just *how* – would Bima undress? In the data we see the passive markers, for example, in all of their tense-specific proliferation. What process would lead speakers of a language to abandon *all* of these morphemes? Crucially, Keo, Rongga and Ngadha indicate no phonetic remnant of these markers, such that, to take a hypothetical example, verbs of high transitivity disproportionately begin with alveolar consonants due to colouring from the erosion of *di-*.

6. Applying the Hypothesis

In sum, Riau Indonesian and the East Timorese languages can be seen as consonant with my hypothesis that languages anomalously less grammatically complex than their sisters have extensive non-native acquisition in their histories. The Flores languages, however, remain quizzical.

There are two possible verdicts on these languages as they relate to my hypothesis.

One is to treat them as contradictions. The idea would be that only a weak version of my stipulation holds: that radical loss of complexity is *usually* due to extensive non-native acquisition, but that it can also happen naturally.

There are problems with that choice, however. One is that, as I have mentioned, there is no known mechanism for this kind of loss: we can describe it, but not explain it.

Another problem with treating Keo, Rongga and Ngadha as evidence that my stipulation must make room for chance exceptions is that at this point, the anomaly appears to be so extremely rare. At this point it would appear that the only languages on the planet Earth where this bizarre process has occurred are a few spoken on the one tiny island of Flores – and even then, amidst others of the island like Sika that have evolved normally. If all but a handful out of 6000 languages conform to a pattern, then it is scientifically appropriate to propose that this handful can be explained in such a way that brings them under the pattern.

As such, a second verdict on these languages is one which I at present prefer. Given that elsewhere, this degree of loss of complexity correlates so tightly with large-scale non-native acquisition, and that this is indeed *the only logical process known or logically constructed for such loss*, we can reasonably propose that the typological contrast between these Flores languages and other Austronesian languages is itself evidence of a break in transmission in the past.

This is, indeed, a logical upending of how linguists have usually reasoned on the relationship between grammatical complexity and sociohistory. Typically, we observe peculiar simplicity and then chart non-native acquisition in the language's history and argue that the latter caused the former. Here, I propose that we might venture, given the wealth of cases in which unusual simplicity can be responsibly linked to transmissional rupture, that a case of unusual simplicity indicates that such rupture occurred and beckons investigation.

Perhaps further descriptive and comparative work will reveal lexical or grammatical interference in these languages on the order of the Ambonese imprint upon the languages of Timor. Folkloric data might also give useful indications.[4] It would also be useful for taxonomic work on these languages to reveal whether they all trace to a single analytic ancestor, since it may be that our anomaly properly comprises a single language that later split into several. It would be especially unlikely that just a single language contravened the tenet of language change I stipulate.

To review, the justification for such an investigation is that languages like English, the Romance languages, Swahili, Kituba, Fanakalo, Mandarin Chinese, Persian, and Indonesian are all markedly less complex than their ancestral languages (cf. Goyette 2000 on Romance), and all are languages associated with the imposition of languages upon peoples amidst

4. My speculation (in Brockman 2006: 71–73) that contact with the recently discovered *Homo floresiensis* may have been the culprit is not intended as a scientific argument; at this point there is not enough paleontological or archaeological data to construct a substantial, refutable account. However, the typological anomaly of these languages is fascinatingly peculiar enough to make it worthwhile to examine all possibilities.

vast human migrations. As geologists treat cracked quartz as a sign of volcanic eruptions in the past, linguists might treat the strange simplicity of Keo, Rongga, and Ngadha as evidence of social disruption in the past.

Abbreviations

1	first person
2	second person
3	third person
ACC	accusative
AGR	agreement
APP	applicative
CAUS	causative
CLAS	classifier
CONJ	conjuction
ERG	ergative
EVID	evidential
EX	exclusive
FUT	future
IMP	imperative
INST	instrument
ITER	iterative
LOC	locative
NEG	negative
OBJ	object
OO	object oriented
P	plural
PASS	passive
PERF	perfect
POSS	possessive
QUOT	quotative
REC	reciprocal
REV	reversive
S	singular
V	verbal

References

Arka, W. 2006. A note on numerals and classifiers in Rongga. Paper delivered at The 10th International Conference on Austronesian Linguistics (10-ICAL), January 2006, Palawan, Philippines.

Arndt, P.P. 1933. *Grammatik der Ngad'a Sprache* [Verhandelingen: Koninklijk Bataviaasche Genootschap van Kunsten en Wetenschappen 72:3]. Bandung: A.C. Nix.

Baird, L. 2002. A Grammar of Kéo: A Language of East Nusantara. PhD dissertation, Australian National University.

Brockman, J. (ed.) 2005. *What We Believe But Cannot Prove*. New York: HarperPerennial.

Dahl, Ö. 2004. *The Growth and Maintenance of Linguistic Complexity*. Amsterdam: John Benjamins.

Donohue, M. 1999. *A Grammar of Tukang Besi*. Berlin: Mouton de Gruyter.

Drabbe, P. 1963. *Drie Asmat-dialecten* [Verhandelingen van het Koninklijk Instituut voor Taal-, Land- en Volkenkunde, Deel 42]. S'Gravenhage: Martinus Nijhoff.

Durie, M. 1985. *A Grammar of Acehnese on the Basis of a Dialect of North Aceh*. Dordrecht: Foris.

Foley, W.A. 1986. *The Papuan Languages of New Guinea*. Cambridge: Cambridge University Press.

Gil, D. 2001. Creoles, complexity, and Riau Indonesian. *Linguistic Typology* 5: 325–371.

Gil, D. 2002. The prefixes *di-* and *N-* in Malay/Indonesian dialects. In *The History and Typology of Western Austronesian Voice Systems*, F. Wouk & M. Ross (eds), 241–283. Canberra: Pacific Linguistics.

Gil, D. 2005. Isolating-Monocategorial-Associational Language. In *Categorization in Cognitive Science*, H. Cohen & C. Lefebvre (eds), 348–380. Amsterdam: Elsevier.

Gil, D. To appear. Creoles, complexity and associational semantics. In *Deconstructing Creole: New Horizons in Language Creation*, U. Ansaldo and S.J. Matthews (eds). Amsterdam: John Benjamins.

Goyette, S. 2000. From Latin to early Romance: A case of partial creolization? In *Language Change and Language Contact in Pidgins and Creoles*, J. McWhorter (ed.), 103–131. Amsterdam: John Benjamins.

Grinevald, C. & Seifart, F. 2004. Noun classes in African and Amazonian languages: towards a comparison. *Linguistic Typology* 8: 243–285.

Hajek, J. & Tilman, A.V. 2001. *East Timor Phrasebook*. Victoria, Australia: Lonely Planet Publications.

Hansson, I.-L. 1976. What we think we know about Akha grammar. Paper presented at the Ninth International Conference on Sino-Tibetan Languages and Linguistics, Copenhagen.

Hull, G. 1998. The basic lexical affinities of Timor's Austronesian languages: A preliminary investigation. *Studies in the Languages and Cultures of East Timor* 1: 97–174.

Hull, G. 2001a. *Baikenu Language Manual for the Oecussi-Ambeno Enclave (East Timor)*. Winston Hills, Australia: Sebastião Aparício da Silva Project.

Hull, G. 2001b. A morphological overview of the Timoric Sprachbund. *Studies in Languages and Cultures of East Timor* 4: 98–205.

Hyman, L. Forthcoming. How to become a Kwa verb. *Journal of West African Languages*.

Jonker, J.C.G. 1915. *Rottineesche Spraakkunst*. Leiden: E. J. Brill.

Kihm, A. 2003. Inflectional categories in creole languages. In *Phonology and Morphology of Creole Languages*, I. Plag (ed.), 333–363. Tübingen: Niemeyer.

Klamer, M. 1998. *A Grammar of Kambera*. Berlin: Mouton de Gruyter.

Lee-Smith, M.W. 1996. The Tangwang language. In *Atlas of Languages of Intercultural Communication in the Pacific, Asia, and the Americas* (Volume II.2), S.A. Wurm, P. Mühlhäusler & D.T. Tryon (eds), 875–882. Berlin: Mouton de Gruyter.

Lee-Smith, M.W. & Wurm, S.A. 1996. The Wutun language. In *Atlas of Languages of Intercultural Communication in the Pacific, Asia, and the Americas* (Volume II.2), S.A. Wurm, P. Mühlhäusler & D.T. Tryon (eds), 883–897. Berlin: Mouton de Gruyter.

Lewis, E.D. & Grimes, C.E. 1995. Sika. In *Comparative Austronesian Dictionary*, Part I, Fascicle I., D.T. Tryon (ed.), 601–609. Berlin: Mouton de Gruyter.

McWhorter, J.H. 2001a. The world's simplest grammars are creole grammars. *Linguistic Typology* 5: 125–166.

McWhorter, J.H. 2001b. What people ask David Gil and why: Rejoinder to the replies. *Linguistic Typology* 5: 388–412.

McWhorter, J.H. 2002. What happened to English? *Diachronica* 19: 217–272.

McWhorter, J.H. 2007. *Language Interrupted: Signs of Non-native Acquisition in Standard Language Grammars*. New York: Oxford University Press.

McWhorter, J.H. Forthcoming. What else happened to English? A brief for the Celtic hypothesis. *English Language and Linguistics*.

Moussay, G. 1981. *La langue minangkabau*. Paris: Archipel.

Mufwene, S.S. 1997. Kituba. In *Contact Languages: A Wider Perspective*, S.J. Thomason (ed.), 173–208. Amsterdam: John Benjamins.

Müller-Gotama, F. 2001. *Sundanese*. Munich: Lincom Europa.

Prentice, D.J. 1990. Malay (Indonesian and Malaysian). In *The World's Major Languages*, B. Comrie (ed.), 913–935. New York: Oxford University Press.

Schwegler, A. 1998. El Palenquero. *América Negra: Panorámica Actual de los Estudios Lingüísticos sobre Variedades Hispanas, Portuguesas y Criollas*, M. Perl & A. Schwegler (eds), 219–291. Frankfurt: Vervuert.

Smith-Hefner, N.J. 1988. Cara Tengger: Notes on a non-standard dialect of Javanese. In *Studies in Austronesian Linguistics*, R. McGinn (ed.), 203–233. Athens OH: Ohio University Center for International Studies.

Van Klinken, C.L. 1999. *A Grammar of the Fehan Dialect of Tetun*. Canberra: Pacific Linguistics.

Williams-Van Klinken, C., Hajek, J. & Nordlinger, R. 2002. *Tetun Dili: A Grammar of an East Timorese Language*. Canberra: Pacific Linguistics.

Morphological complexity as a parameter of linguistic typology

Hungarian as a contact language

Casper de Groot
University of Amsterdam

The paper builds on studies on Hungarian spoken outside Hungary (Fenyvesi (ed.) 2005), which show a change from synthetic to analytic expression in Hungarian in contact. It argues that a parameter of morphological complexity is helpful to account for most morphological changes. With one exception the changes follow the strategy of replicating use patterns (Heine & Kuteva 2005). Other changes arise by implication of a different typological system adopted by the new varieties of Hungarian (De Groot 2005a). A detailed comparison between Hungarian inside and outside Hungary in terms of linguistic complexity (Dahl 2004) confirm to the idea that languages in contact become linguistically more complex. The paper furthermore discusses the interaction between typology, language change by contact, and complexity.

1. Introduction and the aim of the paper

Studies on Hungarian spoken outside Hungary (Austria, Slovakia, Ukraine, Romania, Vojvodina, Prekmurje, United States and Australia as presented in Fenyvesi (ed.) 2005) reveal interesting information on the change of Hungarian as a minority language.[1] Most studies in that volume present data elicited on the basis of the questionnaire which was developed in the "Sociolinguistics of Hungarian Outside Hungary Project" carried out in the early 1990s. Where available, the data was completed by data from earlier studies. In the questions two sentences or linguistic constructions were contrasted. The task of the informant was to tell which sentence or construction in his or her opinion sounded more natural, or in other cases whether sentences or constructions should be considered good or bad. Consider e.g., number 514 of the questionnaire (Fenyvesi (ed.) 2005: 403).

> "Out of the following pairs of sentences, circle the number corresponding to the sentence you consider to be more natural sounding."

1. Vojvodina and Prekmurje are two different regions of the former Yugoslavia. These two regions are now found in separate states namely in Serbia and Montenegro and in Slovenia.

514 (1) *Tanító néni, fáj a fejem. Kimehetek?*
(2) *Tanító néni, fáj a fejem. Ki tudok menni?*
'Miss, I have a headache. May I go out?'

Most of the differences – if not all – between standard Hungarian (HH) and Hungarian outside Hungary (HO) can be explained in terms of language contact where HO takes over features from adjacent languages, which all happen to be Indo-European. In De Groot (2005a) I showed that a number of the differences observed in all varieties of Hungarian outside Hungary follow linguistic universals and implicational hierarchies and that the co-occurrences of changes can actually be explained in terms of universals or hierarchies.

However, since typological literature does not offer explanations for the systematic change from synthetic expression in HH to analytic expressions in all the varieties of HO, I suggested in de Groot (2005a) that a typological parameter of morphological complexity could account for the differences in the various morphological components. The transition from HH, which is morphologically relatively complex, to HO, which is morphologically less complex, would account for why speakers of the latter favour less complex words in cases where speakers of HH would use morphologically more complex words. Further typological research on a large sample of languages in the field of language variation should be carried out to see whether the parameter of morphological complexity is a typological parameter indeed. This paper will not address this issue. It will, however, investigate the status of the notion of complexity in the parameter suggested, i.e., to what extent the notion of complexity is a relevant metric within the domain of morphology and how does it relate to the grammatical system as a whole. The discussion will be based on largely the same set of data presented in Fenyvesi (ed.) (2005) and my analyses thereof in that volume.

The paper is organized in the following way. Section 2 presents a number of distinctions relevant to the discussion of the data related to complexity. These distinctions and the general background are based on Dahl (2004). Section 3 takes the data presented in De Groot (2005a) as its input. For each morphological category, more detailed grammatical and typological analyses are given for differences which hold between forms used in HH and HO. The analyses enable better insights in why and how alternative expressions in HO come into existence and use. Furthermore, the differences will be discussed in terms of complexity as presented in Section 2. Section 4 is devoted to a syntactic pattern that arises in one of the new analytic expressions. Section 5 is dedicated to language contact and Section 6 to linguistic typology. Finally, Section 7 presents the general conclusions.

2. Linguistic complexity

For discussing the data, I adopt a number of notions from the work by Östen Dahl (2004) on linguistic complexity. From a theoretical point of view, these notions offer

a rather neutral or descriptive framework, which enables one to formulate clear statements about the degree or type of complexity involved in the transition of HH to HO. The description of the notions is necessarily brief; for a detailed discussion of the notions, I refer to Dahl (2004).

First of all there is the overall notion of *system complexity*. Dahl regards the set of messages that can be expressed in the language under study as given and considers the complexity of the language seen as a system which maps these messages to expressions. One could also ask for the complexity of expressions in the language rather than that of the system. This brings us to the notion of *structural complexity*, a general term for complexity measures that pertain to the structure of expressions at some level of description. A further notion is the *length of derivational history,* that is, the number of steps necessary to generate the expression in a formal system. *Grammatical regulations* (usually called "rules") often involve conceptual distinctions, e.g., "animacy". *Conceptual complexity* is directly correlated to the length of the definition of a concept. In feature-based theories of semantics, a complex meaning would be one that corresponds to a large number of semantic features.

Further important notions are that of *choice structure* versus *output structure*. These notions are closely related to two "levels of grammar", namely tectogrammatics and phenogrammatics:

– tectogrammatics – the study of grammatical structure in itself.
– phenogrammatics – how grammatical structure is represented in terms of expressions.

For instance, the expressions *two pound butter, two pounds butter* and *two pounds of butter* might be claimed to be different ways of realizing the same grammatical construction and thus differ only with respect to their phenogrammatics. A *choice point* should correspond to a decision made by the speaker. Choice points may be *free* or *bound*. Free choice points relate to the tectogrammatics, where bound choice points are part of the phenogrammatics.

Linguistic pattern is used in a theoretically neutral way. Patterns, referred to as *linguistic objects*, may be of simple or complex types. Simple patterns would be words or morphemes. Complex patterns may consist of fixed parts only, such as a set of phrases. A slot in a pattern is a choice point. The term *construction* may be used for schematic patterns and *auxiliary pattern* for an element that helps build up an expression of another pattern but does not constitute an independent communicative choice. Auxiliary patterns may develop an identity of their own, by appearing in many different constructions. They would still belong to the phenogrammatics, however, representing bound rather than free choices on the part of the speaker.

Another important notion is that of *linearity*. If linearity does not play a role in the grammar of a language, i.e., there is a system of unrestricted concatenation, the language has zero phenogrammatics, and hence zero complexity in this respect. The phenogrammatical complexity of a language is then the extent to which it (or rather the grammar) deviates from a system of unrestricted concatenation. I will take this

statement as the perspective from which I will discuss the differences between the data from HH and HO. Different types of deviations can be distinguished which I will refer to during the various discussions.

The examples to be discussed in this paper concerning different uses in HH opposed to HO all show a difference in terms of synthetic expressions versus analytic expressions. I also raise the question whether there is a degree in phenogrammatical complexity between these synthetic and analytic expressions. Note that the opposite expressions in the comparison have the same context and are meant to convey the same grammatical aspects. Thus, in all cases, there is one choice structure but two output structures.

3. Morphology

Where HH uses morphologically complex synthetic forms, all varieties of HO show a preference for morphologically simplex analytic forms (Table 1). Note that the preference for analytic forms applies in all three major components of morphology, i.e., inflection (modality), derivation (reflexive and causative), and compounding.

Table 1. Synthetic versus analytic expressions in Hungarian inside – outside Hungary.

Section/type	Hungarian inside Hungary (synthetic)	Hungarian outside Hungary (analytic)
3.1 Modality	*Ki-me-**het**-ek?* out-go-MOD-1SG 'May I go out?'	*Ki **tud**-ok men-ni?* out be.able-1SG go-INF 'May I go out?'
3.2 Reflexive	*Szépít-**kez**-ett* beautify-REFL-PAST.3SG.INDEF 'She beautified herself.'	*Szépít-ette **magá-t*** beautify-PAST.3SG.DEF oneself-ACC 'She beautified herself.'
3.3 Causative	*Meg-rajzol-**tat**-ta* ASP-draw-CAUS-PAST.3SG.DEF *a szék-et.* the chair-ACC 'S/he had the chair designed.'	***Hagy**-ta a szék-et* permit-PAST.3SG.DEF the chair-ACC ***rajzol-ni**.* draw-INF 'S/he had the chair designed.'
3.4 Compounding	*tag-létszám* member-number 'number of members'	*tag-ok létszám-a* member-PL number-3SG.POSS 'number of members'
3.5 Multiply derived form	*busz-oz-ás* Bus-VDER-NDER 'bus trip'	*utaz-ás busz-szal* travel-NDER bus-INSTR 'bus trip'

The status of the different analytic expressions in this table is not always the same. The analytic forms may be available in HH as well, although used in a slightly different way. The opposite expressions may also have a different meaning. For a better understanding

of the differences which hold between the synthetic expressions used in HH and the analytic forms preferred in the varieties of HO, I will discuss the data in more detail.

3.1 Modality

Instead of the suffix *-hat/-het* in HH, the free form *tud* 'know, be able' is considered more natural in most varieties of HO.[2] The analytic form also exists in HH, however, used in a more restricted way. Although both expressions relate to modality, they denote different modalities. The difference between the two modalities could be captured in the following terms. The form with *tud* is a "participant-oriented modality", which concerns the relation between a participant in a situation and the potential realization of that situation. The form with *-hat/-het* is a "situation-oriented modality", which concerns the objective assessment of the actuality status of the situation. Compare the following two forms from HH:

(1) a. *Meg tud-om csinál-ni.*
ASP be.able-1SG make-INF
'I can do it.' (I am able to do it)
b. *Meg-csinál-hat-om.*
ASP-make-MOD-1SG
'I can do it.' (Nothing prevents me from doing it.)

What we see in the varieties of Hungarian outside Hungary is that the construction with *tud* takes over the function of situation-oriented modality as in (1b). This kind of change is found more often in languages. English *will* had the function of "person-oriented modality in Old English" which developed via the function of "situation-oriented modality" into the "future tense" in Modern English (Foley & Van Valin 1984: 217). In fact, the change of grammatical operators, such as aspect, tense, and modality, into operators with wider scope seems to be universal (De Groot 1995: 42). The change of the use of *tud* in HO for both "person-oriented modality" and "situation-oriented modality" is thus in line with the universal picture of the historical change of operators in natural languages. The wider use of the verb *tud* 'know, be able' as a modal verb in HO contributes to the process in which the lexical meaning of the verb is fading. Numerous examples in the languages of the world illustrate the development of lexical verbs into auxiliaries where the lexical verbs gradually loose their meaning. Again, the change in the use of verbs in periphrastic constructions attested in HO conforms to the typological development (cf. Bybee, Perkins & Pagliuca 1994; Olbertz 1998).

---

2. The choice of the suffix, having a back or front vowel, is determined by vowel harmony. The suffix will take the form of *-hat* when attached to a verbal stem with a back vowel, such as *lát-hat* (see-MOD) 'may see'. Following the convention of Hungarian grammarians, I will mention all different forms when referring to just one suffix.

Let us now see whether there is a difference in phenogrammatical complexity between the synthetic and analytic expressions of modality, here repeated as (2).

(2)

HH	HO		
Ki-me-het-ek?	*Ki*	*tud-ok*	*men-ni?*
out-go-MOD-1SG	out	be.able-1SG	go-INF
'May I go out?'	'May I go out?'		

In order to arrive at the expressions, the speaker will use the following types of lexical and grammatical input.

(3)

Lexical element:	verb [*kime-/kimen-/kimegy-*] 'go out'
Subject:	first person singular
Grammatical operators:	modality; interrogativity
Pragmatic context:	all new information

Ki- 'out' is a preverbal element, which is separable from the stem. In this particular pragmatic context (see example 514 of the questionnaire in Section 1) the preverbal element *ki* 'out' takes the position immediately preceding the stem in the synthetic expression and the one immediately preceding the modal verb in the analytic expression. First person singular subject is expressed on the finite verb by means of a referential suffix. The grammatical operator of modality takes the form of a bound morpheme in the synthetic expression and the form of a verb in the other. In the latter case, the lexical element will take the form of the infinitive, which could be taken as an extra grammatical operation. The operator of interrogativity (YES/NO) triggers a sentence intonation contour with high pitch on the pen-ultimate syllable and deep fall on the last syllable. The synthetic expression consists of four morphemes united in one word (*Ki-me-het-ek* [out-go-MOD-1SG]), whereas the analytic expression consists of five morphemes spread over three words, i.e., free phonological forms (*Ki tud-ok men-ni* [out be.able-1SG go-INF]). The concatenation of elements in both constructions is equally restricted, i.e., there is, given the context, no free variation. In the case of the analytic expression, permutations are possible but constrained by pragmatics; the different permutations correspond to different information structures.

Where does this bring us? There is not much difference between the two expressions at the first sight. Linearity is equally rigid and in terms of the number of morphosyntactic operations, the analytic expression requires one more operation than the synthetic expression: the formation of the infinitive. When we have a look at the patterns, we could measure complexity in terms of morphosyntactic units: morphemes within the domain of words and free phonological forms (words) within the domain of clauses. From this point of view there is a clear difference between the expressions; the synthetic expression contains more morphemes, whereas the analytic expression contains more words. Another aspect of patterns is that HO uses just one form, the analytic form, to express two different modalities. In terms of Dahl's (2004: 120f) *pattern regulations*, this instantiates a decrease in system complexity, because just one construction expresses two different modalities.

By way of summarizing the differences, consider (4), in which the synthetic expressions seem overall a bit more complex than the analytic ones.

(4)

	HH	HO
	Synthetic expression	Analytic expression
Word structure	more complex	less complex
Clause structure	less complex	more complex
Linearity	complex	complex
Patterns	more complex	less complex

In addition to this comparison there is a difference at the level of the tectogrammatics. The choice structure for the analytic expression is richer than that for the synthetic expression. If a speaker wishes, for instance, to emphasize the modality or the action as a pragmatic focus in a contrastive setting, this would be possible in HO with analytic expressions, but not in HH. The expressions in HO would be (5): note the position of the preverbal element *ki* 'out' opposed to the expression in (2) as wells as the order of constituents.

(5) HO

a. *Tud-ok* *ki-men-ni?*
be.able-1SG out-go-INF
'MAY I go out?'

b. *Ki-men-ni* *tud-ok?*
Out-go-INF be.able-1SG
'May I GO OUT?'

From this it follows that a speaker of HO has in this respect more choice points, hence greater tectogrammatical complexity is involved.

On the basis of the data discussed here and their degree in various types of complexity, we may conclude that HH with synthetic expressions is morphologically more complex whereas HO is more complex in the domain of syntax. The same conclusion can *mutatis mutandis* be drawn on the basis of the other data of Table 1. Therefore, I will not discuss similar oppositions between expression from HH and HO in the following sections in the same detail.

3.2 Reflexive

Instead of the derived verbal reflexive with the suffix *-koz(ik)/-kez(ik)/-köz(ik)* in HH, the transitive verb form is more preferred in the different varieties of HO. The example from the questionnaire illustrating this in Table 1 is here repeated as (6).

(6) HH *Szépít-kez-ett*
beautify-REFL-PAST.3SG.INDEF
'She beautified herself.'

HO *Szépít-ette* *magá-t*
beautify-PAST.3SG.DEF oneself-ACC
'She beautified herself.'

Hungarian has a very rich system of derivational morphology, where a great number of the derivational relations involve a change in the valency of the verb (cf. De Groot 1989: ch.5). Among them, there are various examples of detransitivization, one of which is the verbal reflexive with the schematic pattern in (7):

(7) Second argument reduction
Input: Verb (x1) (x2)
Output: Verb-sx (x1)

The detransitivized pattern is used if the action performed by the agent applies to oneself. When we apply these patterns to the verb *borotvál* 'shave', we get the following picture, where (8b) is considered ungrammatical in standard Hungarian. Also note that the overt expression of an object in (8c) is ungrammatical, which clearly indicates that verbal reflexives are intransitives. Compare:

(8) a. *A borbély borotválja Feri-t.*
the barber shave Feri-ACC
'The barber shaves Feri.'
b. ?**A borbély borotválja magá-t.*
the barber shave himself-ACC
'The barber shaves himself.'
c. *A borbély borotvál-koz-ik (*magá-t)*
the barber shave-DER-3SG (himself-ACC)
'The barber shaves himself.'

It is relevant that the productivity of the formation of verbal reflexives is rather low. First of all it is very difficult, if not impossible, to give a clear description of the class of verbs which can serve as the input to the derivational rule, and second, many of the verbal reflexives have been lexicalized (De Groot 1989: 142). That may explain why *borotválkozik* 'shave oneself' does not allow a transitive counterpart with a reflexive object (cf. (8b)). Strictly speaking, *borotválja magát* is not ungrammatical nowadays in HH, but it is associated with intimate lady shaving. In other words, it has a newly created idiomatic interpretation, which does not naturally follow from the transitive reflexive form – the reason why example (8b) is marked as "ungrammatical".

The class of transitive verbs that does not allow the formation of verbal reflexives remains transitive when used in a reflexive way. In those cases the object is specified by a reflexive pronoun:

(9) HH
a. *Nézi magá-t a tükör-ben.*
See.3SG himself-ACC the mirror-INES
'He sees himself in the mirror.'
b. *Megvágtam magam-at.*
ASP.cut.PAST.1SG myself-ACC
'I cut myself.'

In other words, one class of verbs takes the derived intransitive pattern whereas another class of verbs takes the transitive pattern to express reflexivity in HH. In HO there is a preference to use just one pattern, namely the transitive pattern. This constitutes an example of *pattern regulation*; it instantiates a decrease in system complexity because just one pattern expresses reflexivity for all classes of verbs. Since many of the verbal reflexives are lexicalized, the use of the transitive form of those verbs in HO could be considered instances of "back-formation". Under this view, the use of the transitive pattern in cases such as *szépíti magát* 'beautify oneself' constitutes one more step in the derivational history, namely that of back formation, which would then count as an increase of complexity.

In contrast to the expressions of reflexivity, where there are two patterns, an intransitive and a transitive pattern, there are other examples of derived intransitives which lack a transitive pattern as an alternative for other classes of verbs. In those cases, HO does not employ transitive patterns but uses the intransitive patterns like in HH. Consider the following examples where (10a) shows the basic transitive verb and (10b) the derived intransitive, an instance of "first argument (agent) reduction". HH, which lacks a passive construction comparable to English *The door has been closed by John*, does not offer an alternative transitive pattern in this case. Examples such as (10b) are used in both HH and HO.

(10) HH

a. *János bezárja az ajtó-t.*
John close the door-ACC
'John closes the door.'

b. *Az ajto bezár-ód-ik (*János által)*
the door close-DER-3SG (John by)
'The door closes.'

There is no generalization possible in terms of derived intransitives in HH taking the transitive pattern in HO; this would be an instance of a great decrease in complexity. HO uses the transitive pattern only in those cases where the pattern is available in HH.

In addition to the observations already made about complexity in the former section and this one, I would like to point at a difference between intransitive verbal reflexives and transitive forms which may express reflexivity. When the intransitive form is used, the specification of the subject suffices to express that the action performed applies to oneself. When using the transitive form, both the position of the subject and object must be specified in such a way that they refer to the same participant, for instance by co-indexation. Furthermore there will be some kind of agreement between the subject and the object. Compare the differences in (11).

(11)

HH	HO	
Szépít-kez-em.	*Szépít-em*	*maga-m-at.*
beautify-REFL-1SG	beautify-1SG	self-1SG-ACC
'I beautify myself.'	'I beautify myself.'	

We may conclude here that the analytic expression in HO is much more complex in terms of operations to achieve the appropriate expression.

3.3 Causative

Instead of the derived causative with the suffix *-tat/-tet* in HH, the periphrastic construction with the verb *hagy* is considered more natural in HO. The example from the questionnaire in Table 1 is here repeated as (12).

(12) HH *Meg-rajzol-tat-ta a szék-et.*
ASP-draw-CAUS-PAST.3SG.DEF the chair-ACC
'S/he had the chair designed.'
HO *Hagy-ta a szék-et rajzol-ni.*
permit-PAST.3SG.DEF the chair-ACC draw-INF
'S/he had the chair designed.'

The verb *hagy* in Hungarian is a lexical verb with the meaning of 'let, leave, allow, permit'. The vast majority of the use of *hagy* in HH is in the permissive sense. However, the causative application also arises, as can be seen in example (13c).

(13) HH
a. *Hagy-ta magá-t megcsókol-ni.*
allow-PAST.3SG herself-ACC kiss-INF
'She allowed her to be kissed. / She let him kiss her.'
b. *Hagy-juk a múlt-at.*
let-1PL the past-ACC
'Let bygones be bygones.'
c. *Hagy-ja a terv-et fejlőd-ni.*
let-3SG the plan-ACC develop-INF
'(S)he let design a plan.'

The preference of using the periphrastic causative construction in HO over the synthetic construction could be considered an overextension of the periphrastic form. The wider use of *hagy* contributes to the process of auxiliarization and is comparable to the fate of *tud* 'know, be able' as well as to many lexical verbs which are used in periphrastic constructions, namely that they gradually loose their lexical meaning. From the point of view of complexity, the use of just one analytic pattern expressing both a permissive and a causative, instantiates another case of a decrease in system complexity. On the other hand, the predicate structure in the analytic expression is much more complex than in the synthetic one. The verbal element *hagy* 'let' in causative constructions cannot be taken just as a supportive element functioning as a portmanteau for tense and person markers. Rather, the introduction of the element *hagy* 'let' must be the result of a causative formation rule. One may wonder whether the element *hagy* 'let' will keep its own predicate structure or whether it will combine with the lexical verb into some kind of complex predicate. The causative formation in HH is a rule which implies argument

extension: the extra argument has the function of Agent/Causer. In HO two verbal predicates are combined into one frame, where one predicate fills the patient argument position of *hagy* 'let'. The rules are schematically described in (14) and (15):

(14) HH
Input: *rajzol* (Agent) (Patient)
Output: *rajzoltat* (Agent/Causer) (Agent/Causee) (Patient)

(15) HO
Input: *rajzol* (Agent) (Patient)
hagy (Agent) (Patient)
Output: *hagy* (Agent/Causer)
(*rajzol* (Agent/Causee) (Patient))

I conclude that the output of the causative formation in (15) is more complex than the one in (14).

3.4 Compounding

Instead of using compounds consisting of two nouns as in HH, there is a strong preference for the HO varieties to use disjunct expressions. The disjunct expressions come in different forms. There are two basic strategies: (a) the modifying noun, i.e., the first noun in the compound receives an attributive marker, and (b) the creation of a possessive construction. The examples in (16) are illustrations of the first strategy.

(16)

	HH	HO	
a.	*lég-tér*	*légi*	*tér*
	air-space	air-ATTR	space
	'air-space'	'air-space'	
b.	*virág-láda*	*virág-os*	*doboz*
	flower-box	flower-ATTR	box
	'flower box'	'flower box'	

The two expressions with the attributive markers in (16) count as non-standard or even ungrammatical in HH. Constructions with the attributive markers themselves, however, do occur very frequently in Hungarian, as for instance in (17):

(17) HH
a. *a* *tavasz-i* *fesztivál*
the spring-ATTR festival
'the spring festival'
b. *az* *erkély-es* ház
the balcony-ATTR house
'the balconied house'

Moreover, in other constructions than those in (16), the attributive forms *légi* and *virágos* may arise in HH, as in (18). For reasons I do not know, expressions such as

those of HO in (16) are not conceived as grammatical in standard Hungarian, whereas those in (18) are.

(18) HH
a. *lég-i* *forgalom*
air-ATTR traffic
'air traffic'
b. *virág-os* *függöny*
flower-ATTR curtain
'flowered curtain'

The following example illustrates that compounds which consist of an adjective and a noun in HH may also have disjunctive counterparts in HO. Interestingly, the adjectival form – here the present participle *szolgáltató* – which does not allow the attributive marker *-i*, has been substituted by the related form, the nominalization *szolgáltatás*, which does allow the marker *-i*:

(19)	HH		HO	
	szolgáltatóház		*szolgáltatási*	*ház*
	$[[[\text{szolgáltat}]_V\text{-ó}]_A\text{-}[\text{ház}]_N]_N$		$[[[\text{szolgáltat}]_V\text{-ás}]_N\text{-i}]_V$	ház
	serve-DER-house		serve-DER-ATTR	house
	'service house'		'service house'	

Note that a similar opposition can be found in HH as well, as can be seen in example (20). This indicates that the preferred expression in HO does not take a newly created pattern but is based on a pattern available in HH.

(20) HH
bevásárló-központ vs. *vásárlás-i* *láz*
shopping-centre shopping-ATTR fever
'shopping centre' 'shopping boom'

At this point, it can be concluded that the analytic constructions in strategy (a) (the modifying noun), which are favoured in HO, can all be found in HH as well. The meaning, however, can be different in HO and HH. The patterns could act as a model for the use of the analytic forms. Instead of using two patterns, i.e., a compound and an attributive noun construction as in HH, HO uses just one pattern, hence another case of a decrease in system complexity.

One remark on the expression with the attributive marker *-i*. I have argued elsewhere that expressions with the marker *-i* are based on a morphosyntactic template, readily made available in the grammar of Hungarian. The template has the form of [[X]-i N], where X accommodates other elements than adjectives such as nouns, but also postpositions or even a postpositional phrases (see De Groot (2005b) for a full specification and De Groot (1990) for typological principles underlying the patterns). In Dahl's (2004) terminology, this template could be considered an *auxiliary pattern*

that helps build up an expression of another pattern but does not constitute an independent communicative choice. In this case, the output is a phrase with the function of an attribute. Such templates or patterns may be ideal vehicles for certain transitions from HH to HO. It does not seem, however, that the pattern is used extensively in HO. I return to this matter in Section 4.

The second strategy as an alternative disjunct expression for a compound is the possessive construction. Compare the different forms in (21).

(21)

	HH	HO	
a.	*tag-létszám*	*tag-ok*	*létszám-a*
	member-number	member-PL	number-3SG.POSS
	'number of members'	'number of members'	
b.	*nő-nap*	*nő-k*	*nap-ja*
	woman-day	woman-PL	day-3SG.POSS
	'Woman's Day'	'Women's Day'	

The two disjunct expressions in (21) with the meanings 'number of members' and 'Woman's Day', respectively, are not used in HH. However, the disjunct structure used in HO is an existing structure in HH as (22) illustrates.

(22) HH

a	*tag-ok*	*egyesület-i*	*nyakkendő-je*
the	member-PL	club-ATTR	tie-3SG.POSS

'the club-tie of the members'

It is of particular interest that nouns which are used to form compounds in Hungarian cannot take the plural form, unlike e.g., Dutch, contrasted in (23).

(23)

Dutch	Hungarian	
huizenblok	*háztömb*	**házaktömb*
[[huiz-en]-[blok]]	[[ház]-[tömb]]	[[ház-ak]-[tömb]]
house-PL-block	house-block	house-PL-block
'block of houses'	'block of houses'	

The difference between the types of compounds as in Dutch opposed to Hungarian can be explained on the basis of the types of nouns the languages have. An NP headed by a *bare* noun must refer to a single individual in Dutch: when reference is made to more than one individual object, the noun must be suffixed with a plural marker. Rijkhoff (2002) calls such nouns (typically found in the Indo-European languages) *singular object nouns*. Hungarian, on the other hand, has nouns that denote a set of individuals. These are called *set nouns*. A set can have any cardinality: it may contain just one individual or it may consist of more individuals. A property of *set noun* languages is that they lack a plural but instead they may have a collective form (Rijkhoff 2002: 51). The marker *-k* in Hungarian is therefore considered a collective marker rather than a plural marker. De Groot (2005a) argues that, due to language contact, HO develops

from a *set noun*-type language into a *singular object noun*-type language. This accounts for several phenomena in HO because they are properties of *singular object noun* languages: the plural marking of nouns after a numeral, the use of plural form of the noun in certain other cases, and plural agreement with the verb in a number of cases. This insight might tell us something about the preference in HO to use the disjunct form such as *tagok létszáma* (members number of) instead of the compound *tag-létszám* (member-number). Since nouns in HO denote more and more single entities instead of sets, the urge of using the plural form of the first noun in compounds of the type [member-s number] forces the speaker to use the disjunct expression which allows the use of the plural but which are not allowed by compounds. Evidence from The Hungarian National Corpus supports this view, exhibiting the following matches:[3]

(24)	a.	172 times	taglétszám	[member-number]
	b.	8 times	tagok létszáma	[members number-of]
	c.	0 times	tag létszáma	[member number-of]

The data in (24) shows that in those cases where the compound is split up into two parts, the first member is marked by the plural.[4] No instances were found where the first part of the possessive alternative was used in the singular, i.e., as a set noun. The use of the possessive construction instead of the compound could then also be considered a result of the change of HH, a set noun type language, into HO, a singular object noun type language. Table 2 summarizes the changes from HH to HO, which all follow from the typological change of HH to HO.[5]

Table 2. Change of HH – set noun type – into HO – singular object noun type.

	Hungarian inside Hungary **Set noun type**	**Hungarian outside Hungary** **Singular object noun type**
numeral + N	*két könyv* two book 'two books'	*két könyv-ek* two book-PL 'two books'

(*Continued*)

3. At the time of consulting, the corpus consisted of 187,644,886 words from Hungary, Slovakia, Subcarpathia, Transylvania, Vojvodina, note the sources from both HH and HO. The corpus is available at http://corpus.nytud.hu/mnsz/index_eng.html.

4. The context in which one of the examples is presented reveals that the example originates from a scientific text included in the Hungarian National Corpus discussing morphological differences found in HH and HO. One of the differences mentioned in the article is *taglétszám vs. tagok létszáma*.

5. De Groot (2005a) mentions the first three examples. The compound is here added as a fourth example of language change due to a typological change.

Table 2. Continued.

	Hungarian inside Hungary **Set noun type**	**Hungarian outside Hungary** **Singular object noun type**
Number discord vs. concord	*Három fiú sétál.* three boy walk.3SG 'Three boys walk.' *Mari és Péter sétál.* Mary and Peter walk.3SG 'Mary and Peter walk.'	*Három fiú-k sétál-nak* three boy-PL walk-3PL 'Three boys walk.' *Mari és Péter sétál-nak.* Mary and Peter walk.3PL 'Mary and Peter walk.'
set vs. individual	*János almá-t vett.* John apple-ACC bought 'John bought apples.' *Fáj a láb-am.* ache.3SG the leg-POSS.1SG 'My legs are aching.'	*János almá-k-at vett.* John apple-PL-ACC bought 'John bought apples.' *Fáj-nak a láb-ai-m.* ache.3PL the leg-PL-POSS.1SG 'My legs are aching.'
compound (set) vs. possessive (individual)	*tag-létszám* member-number 'number of members'	*tag-ok létszám-a* member-PL number-3SG.POSS 'number of members'

What does the data in Table 2 tell us about complexity? Firstly, all constructions in HO show more structural complexity than the forms in HH, notably the use of the plural. Secondly, there is a change in *Seinsart* from *set* to *singular object*. It would be interesting to know whether there is a difference in complexity between the concepts of *set* versus *singular object*. If there was a difference, it would be a difference in conceptual complexity. Rijkhoff (2002) does not characterize *set* and *singular object* in terms of different sets of features, parameter setting, or alike. There are, however, some clues which point at *singular object* being conceptually more complex than *set*: firstly, *singular object* is characterized in terms of [numeral + noun + plural], whereas *set* as [numeral + noun], which consists of two elements and not three; and secondly, from a cross-linguistic perspective number marking seems to be the exception rather than the rule (Rijkhoff 2002: 38). On the basis of these two observations concerning the use of the plural, i.e., one more morpheme per construction, and marked from a typological perspective, I suggest that *singular object* as a *Seinsart* is conceptually more complex than *set*.

3.5 Multiply derived forms

The last example in Table 1 presents another relevant observation. This has to do with the length of derivational history. Compare the data in (25), here repeated from Table 1.[6]

6. According to the rules of Hungarian orthography and word segmentation, the geminate in the form *busszal* is presented in the following way: *busz-szal*. Double *szsz* is reduced to *ssz* in actual writing, but not in speech.

(25)	HH	HO	
	buszozás	*utazás*	*busz-szal*
	$[[[busz]_N\text{-}oz]_V\text{-}ás]_N$	$[[utaz]_V\text{-}ás]_N$	busz-szal
	bus-VDER-NDER	travel-NDER	bus-INSTR
	'bus trip'	'bus trip'	

Example (25) illustrates that there is a preference in HO for complex derived words to be replaced by forms consisting of smaller or less complex units. The form in HH is a nominalization of a denominal verb, whereas the form in HO is a nominalization followed by a specifying element marked by the instrumental case. The form in HH contains just one lexical element, but the form in HO contains two. Note that the expression used in HO is not available in standard Hungarian. I will return to this matter in Section 4.

There is more evidence that multiply derived forms in Hungarian are complex indeed. Recursiveness in the formation of words is quite possible in Hungarian, as for instance in the following form, which contains four derivational suffixes:

(26)	$tart_V$	$\text{-}ós_A$	$\text{-}ít_V$	$\text{-}ható_A$	$\text{-}ság_N$
	keep |	durable |	conserve |	conservable |	conservability

However, there are many instances where derivational rules in Hungarian only allow non-derived words as their input:

(27) a. Adjective → One-place inchoative verb
$[meleg]_A$ 'warm' → $[[meleg]_A\text{-}edik]_V$ 'get warm'
$[[tart]_V\text{-}ós]_A$ 'durable' → Ø

b. Two-place agentive verb → One-place process verb
$[húz]_V$ 'draw' → $[[húz]_V\text{-}ódik]_V$ 'draw'
$[[[tart]_V\text{-}ós]_A\text{-}ít]_V$ 'conserve' → Ø

The status of a lexical item in Hungarian as derived may block derivational processes (De Groot 1989: 163). Similar phenomena, i.e., that derived forms behave differently – also in the field of inflectional morphology – are found in other languages as well. This does not imply, however, that certain derived forms are excluded from derivational processes in all cases. An example that illustrates this is (28), in which the derived form *tartós* 'durable' may form the input of a derivational rule in (28b) but not in (28a):

(28) a. Adjective → One-place inchoative verb
$[[tart]_V\text{-}ós]_A$ 'durable' → Ø

b. Adjective → Transitive verb
$[[tart]_V\text{-}ós]_A$ 'durable' → $[[[tart]_V\text{-}ós]_A\text{-}ít]_V$ 'conserve'

Rules which account for the correct derivation of new forms, such as those in (27), will be specified in a certain way to prevent the expectation of derived forms as their input, whereas other rules do not need such specification, as for instance (28b).

From the discussion in this section it can be concluded that derived forms are more complex than non-derived forms, because of their length of derivational history (Dahl 2004: 44). Grammatical rules may also be sensitive to the forms being derived, which could be taken to increase system complexity. In these respects, derivation in HH is more complex than in HO.

4. Syntax

The only syntactic issue I raise here concerns the preference of certain post-nominal modification in the varieties of Hungarian outside Hungary. In example (25) we saw that in HO a syntactic expression arises which does not exist in HH. Another similar expression occurs in HO, which I will discuss in more detail in this section.

In equivalence to the English phrases *the picture on the wall* or *the letter from Mary* standard Hungarian employs non-finite pre-nominal constructions to express the modification of the nominal head (29):

(29) *a fal-on levő kép*
the wall-SUP COP.PRES.PART picture
'the picture on the wall' [lit. 'the on the wall being picture']

Examples such as (29) pattern along with non-finite participle constructions, as in (30):

(30) *az éneklő lány*
the sing.PRES.PART girl
'the singing girl'

Constructions like (29) and (30) can be taken to be non-finite relative constructions (cf. De Groot 1989: 191). They both have finite counterparts, which take the form of finite relative constructions, as shown in (31).

(31) a. *a kép, amely a fal-on van*
the picture REL the wall-SUP COP.3SG
'the picture which is on the wall'
b. *a lány, aki énekel*
the girl REL sing.3SG
'the girl who sings'

When we look at data from the varieties of Hungarian outside Hungary, we see a preference for using another construction which widely differs from the one in standard Hungarian. Instead of a non-finite pre-nominal relative construction, we find a post-nominal adpositional phrase, as in (32).

(32) HO
a kép a fal-on
the picture the wall-SUP
'the picture on the wall'

The difference between the form used in HH (29) and HO (32) may be captured in the following fashion:

(33)

HH	HO
Sentential modifier + Noun	Noun + Non-sentential modifier

Interestingly, both differences, i.e., the application of the copula in HH and the non-application of the copula in HO as well as the order of the constituents, i.e., ModN in HH and NMod in HO, must apply at the same time. The non-application of the copula without the change of the position of the modifier yields an ungrammatical output in HO (34a). The same holds if there is a change in the position of the modifier together with the application of the copula (34b):

(34) HO

a. **a* *fal-on* *kép*
the wall-SUP picture
'the picture on the wall'

b. *a *kép* *a* *fal-on* *levő*
the picture the wall-SUP COP.PRES.PART
'the picture on the wall'

In Section 3.4, I mentioned that Hungarian makes use of an auxiliary pattern as an intermediary for other elements than adjectives to become attributives. One of these candidates is the locative modifier of a noun. The modifier either takes the non-finite form with the present participle of the copula (35a) or the marker *-i* (35b) in those cases where the locative function is expressed by a postposition.

(35) HH

a. *a* *fal* *mögött* *levő* *kép*
the wall behind COP.PRES.PART picture
'the picture behind the wall'

b. *a* *fal* *mögött-i* *kép*
the wall behind-ATTR picture
'the picture behind the wall'

Modifiers specified by a case ending disqualify as the input for the auxiliary pattern; the attributive marker *-i* does not combine with case markers (36).

(36) HH

**a* *fal-on-i* *kép*
the wall-SUP-ATTR picture
'the picture on the wall'

What we see in the varieties of HO is that expressions such as (36) do not occur either. The varieties in HO could have made use of the auxiliary pattern available, but they did not, apparently because they are ungrammatical in HH. Instead they

employed a pattern which in standard Hungarian is highly disfavoured if not available at all.[7]

5. Language change by contact

The way in which Hungarian outside Hungary changes looks very much like what happens in many other instances of language change by language contact as brought forward by Heine and Kuteva (2005). In their model of replicating use patterns, grammatical replication has the effect that the replica language (HO) acquires some new structures (HOx) on the model of another language (contact language). The new structure (HOx) is in most cases not entirely new; rather, it is built on some structure (HHy) that already existed in the replica language, and what replica then achieves is that it transforms (HHy) into (HOx). Many of the changes follow this pattern indeed, as:

(37) a. intransitive verbal reflexives → transitive reflexives
b. synthetic modality → analytic modality
c. synthetic causative → analytic causative
d. compounds → disjunct structures

There is one case which does not follow this pattern, because the new adopted pattern is not available in the replica language, namely post-nominal modification. Here it must be postulated that the new adopted structure is a direct borrowing from the contact language.

The motivation for adopting new patterns in HO cannot always be directly related to the use of certain patterns in the contact languages. Many of the contact languages do have intransitive verbal reflexives, yet the transitive counterpart is favoured in HO. One may argue that compounds are disfavoured in the varieties of HO because most of the contact languages do not have compounds. It has, however, been shown that the typological change from *set* to *singular object* causes the use of disjunct structures where HH employs compounds. Moreover, in the variety of Hungarian spoken in Austria, the compound also disappears, while the formation of compounds in German is very, very productive. Like Dutch, German is a *singular object noun*-type language which allows compounds where the first member is marked by the plural as in *Häuserblock* [[house.PL]block] 'block of houses'. The variety of Hungarian spoken in Austria could also take over this structure, but it does not. Apparently the "decisions" made in the change from HH to HO are the following. Being a *singular object noun*-type language, the formation of a compound of the type [N.PL+N] would be theoretically

7. Post-nominal modification can actually also be found in standard Hungarian, particularly in names of articles or poems, as e.g. M. Radnóti's poem *Levél a hitveshez* 'Letter to the spouse'. Purists dislike them because of their association with German influence.

possible in HO. HH, however, does not offer a use pattern of this type but some other pattern, namely that of a possessive construction of the type [N.PL N.POSS]. German does have the use pattern of [N.PL+N]. Instead of adopting the pattern from German, the HO variety in Austria adopts a use pattern from HH. This reinforces the claim by Heine and Kuteva (2005) as described above, that the new expression is built on some structure that already existed in the replica language.

6. Typology

6.1 Morphological complexity as a parameter

A generally expected view on morphological change, as for instance represented by Hock and Joseph (1996: 183), is the following:

> The fate of morphology from Sanskrit to its modern descendants gives credence to the common belief that languages tend to develop in cycles: from isolating to agglutinating, from agglutinating to inflectional (through amalgamation of different affixes into one), from inflectional to isolating (through sound change and analogy), and so on.

This may be expressed schematically as:

(38) agglutinating → inflectional → isolating

When we consider the varieties of Hungarian outside Hungary, these varieties still contain all morphological properties of an agglutinating language (cf. *szépítette, széket,* and *létszáma* in Table 1). The change from HH to HO then does not follow the chain of gradual changes in (38) as attested in many languages in their individual development, or induced by language contact (cf. Heine and Kuteva 2005: 164). It rather is a change from morphologically complex to morphologically simplex, i.e., from an agglutinating language with highly morphematic words to an agglutinating language with less morphematic words. There are actually some examples of the change of Hungarian as a contact language which follow the development as given in (38). These are e.g., the loss of object agreement and possessive markers in Hungarian in the United States (Fenyvesi 2005). Note, however, that these markers are at the periphery of words.

If we take morphological complexity to be a typological parameter, the following subtypes should be distinguished:

(39)

	Complex	→ Simplex
a.	many morphemes per word	→ few morphemes per word
b.	more lexemes per word	→ fewer lexemes per word

An illustration of (39a) is provided by examples (40) and (41), which show that the forms in HH are more morphematic than the preferred forms in equivalent expressions in HO.

(40)

HH	HO	
rajzol-tat-t-a	*hagy-t-a*	*rajzol-ni*
1 2 3 4	1 2 3	1 2
draw-CAUS-PAST-3SG	let-PAST-3SG	draw-INF
'(S)he had it designed.'	'(S)he had it designed.'	

(41)

HH	HO		
busz-oz-ás	*utaz-ás*	*busz-szal*	
1 2 3	1 2	1	2
bus-VDER-NDER	travel-NDER	bus-INSTR	
'bus trip'	'bus trip'		

An illustration of (39b) is example (42) which shows that compounds in HH, consisting of two lexemes, have counterparts in HO where the two lexemes form the basis of two words.

(42)

HH	HO	
tag-létszám	*tag-ok*	*létszám-a*
1 2	1	1
member-number	member-PL	number-3SG.POSS
'number of members'	'number of members'	

Given the systematic differences between HH and HO, the varieties of Hungarian outside Hungary could be considered to have a morphological profile (cf. Heine and Kuteva 2005: 153) other than that of Standard Hungarian. The profiles are not absolute but relative in the sense that the morphological profile of HO can be characterized as simpler than that of HH. A counter-example to this idea – the use of the plural in HO – will be discussed in the next section.

6.2 Rivalry between parameters

In Section 3.4, we saw that HO uses the plural in various cases where HH uses the bare noun (see Table 2), for instance *két könyv* (HH two book) versus *két könyv-ek* (HO two book-PL). These facts contradict the typological parameter of morphological complexity as presented in the previous section. That parameter predicts that forms in HO will likely have fewer morphemes than their counterparts in HH. In the case of plural it is just the opposite. We must conclude that the parameter which causes the introduction of the plural is more powerful than the parameter of morphological complexity. The stronger parameter here is that of *Seinsart*. The change from HH as a *set noun type* language to HO as a *singular object type* language causes the introduction. The *Seinsart* of a language may be considered a relatively deep property of the language, because it affects a substantial part of the grammar of a language. On the basis of these observations I would like to suggest that *Seinsart* belongs to the tectogrammatics of a language and the morphological complexity to the phenogrammatics. If this is correct, it explains on the one hand why *Seinsart* as a parameter is stronger than morphological

complexity, and on the other hand why exceptions to the parameter of morphological complexity may arise.

6.3 Chicken or egg: Post-nominal modification in HO

As for the one syntactic parameter discussed, I argued in De Groot (2005a) that copula support in HH is sensitive to the hierarchy in (43) (from Hengeveld 1992). The hierarchy tells us that HH employs a copula in relational non-verbal expressions (*John is in the garden*) but not in referential (*John [is] the postman*) or bare (*John [is] young*) non-verbal expressions. The notation // indicates the cut-off point in the hierarchy.

(43) Copula support in non-verbal expressions HH
Finite constructions: Bare > Referential // Relational
Non-finite constructions: Bare > Referential // Relational

The use of copula in the Hungarian varieties outside Hungary shows a different picture. The use of the copula in non-finite embedded relational constructions is no longer favoured. The following opposition holds between HH and HO:

(44) HH *a fal-on levő kép*
the wall-SUP COP.PRES.PART picture
'the picture on the wall' [lit. 'the on the wall being picture']
HO *a kép a fal-on*
the picture the wall-SUP
'the picture on the wall'

The tendency to use constructions without a copula in embedded constructions in HO suggests that the cut-off point for copula support in embedded non-finite constructions in HO moves from before Relational to after Relational. Consider the following hierarchies:

(45) Copula support in HO
Finite constructions: Bare > Referential // Relational
Non-finite constructions: Bare > Referential > Relational //

Typological studies on language variation based on hierarchies clearly indicate that variation is typically found at the cut-off points in hierarchies (Dik 1997). This now also holds for the variation in the use of a copula in HO. In other words, the change in the use of the copula between HH and HO clearly follows typological principles. There is, however, no intrinsic motivation for the cut-off point to move; but when we take the change in the position of the modifiers from pre-nominal to post-nominal as a starting point, then there is. First, non-finite relative constructions in post-nominal positions are rare. Note that relative constructions in languages with basic SOV word order can take two positions in the noun phrase, before or after the nominal head. The position is almost without exception conditioned by the nature of the relative clause, being non-finite or finite. Non-finite relative clauses precede the head, whereas finite relative clauses follow the head. The deletion of the non-finite copula form in post-nominal

constructions avoids the anomaly. A movement of one category to the right on the hierarchy facilitates the deletion. The motivation to use post-nominal modification must come from the contact languages: these languages do not employ pre-nominal non-finite relative clauses. The explanation presented here opposes the view defended in De Groot (2005a) that the change in cut-off point accounts for all aspects, i.e., the deletion of the copula and the change of position of the modifier. The syntactic change is taken to be prior to the deletion of the copula in this paper now.

In terms of complexity, the copula deletion decreases syntactic complexity (structural complexity) but increases morphological complexity if we take zero marking to be more complex than overt marking (cf. Lefebvre 2001; Aboh & Ansaldo 2006). Moreover the move of the cut-off point in the hierarchy of non-finite constructions increases system complexity due to the asymmetry with the hierarchy of finite constructions. However, there seems to be a slight decrease in complexity because the copula deletion may be interpreted as avoiding an anomalous construction and one which is structurally more complex than is typically found in verb-final languages.

7. Conclusions

This paper shows that speakers of Hungarian inside Hungary express their messages differently from speakers of Hungarian outside Hungary. The most striking difference is that where HH uses synthetic expressions, HO uses analytic expressions in many of these cases. The analytic expressions are mostly based on patterns available in HH. In order to accommodate these expressions, the use of certain patterns is extended by way of overgeneralization, or by hosting more than one expression. In terms of linguistic complexity, the variants of HO thus seem less complex than HH. HO is morphologically less complex and it uses fewer patterns for moulding its expressions, which can be taken as a decrease in system complexity. Opposed to the decrease, there is, however, such a substantial increase in system complexity that the variants of HO have to be considered more complex than HH. The increase of complexity is located in several domains: (1) analytic expressions are subject to various rules concerning linearity, (2) periphrastic expressions are mainly built on complex verbs or verbal clusters, and (3) various alternative expressions require extra operations to account for phenomena relating to agreement, co-indexation of referents in reflexive and possessive constructions, or marking dependent relations.

For languages in contact to become linguistically more complex has been observed by several linguists, notably Nichols (1992: 193). She argues that contact among languages foster complexity, in particular in those cases when there is a long-term language contact involving child-language acquisition. The variants of HO belong to this group and the data and analyses presented in this paper confirm the observation that languages in contact become linguistically more complex.

On the relation between typology, language change by contact, and complexity, some interesting conclusions can be drawn. Firstly, HO takes over a deep typological

property of the contact languages, namely the *Seinsart* of *singular object noun* type, which as a type is rather rare among the languages of the world. Secondly, morphological complexity can be considered a relevant parameter in language contact. The parameter mainly relates to surface phenomena, i.e., the number of morphemes and how they are attached to each other. Deeper properties can easily overrule constraints on morphological complexity, such as *Seinsart* which for instance introduces the use of the plural in HO. On the basis of these observations I suggest to consider *Seinsart* as belonging to the tectogrammatics of a language and morphological complexity to the phenogrammatics. Thirdly, due to a change from synthetic to analytic way of expressing, language becomes more open to syntactic rules. Note that HO did not lose many inflectional or derivational rules. The application of modal operators or the formation of causative constructions is still there. The expression, however, is analytic and not synthetic. Fourthly, "pattern" plays an important role as an intermediate factor in language change and system regulation. Under different names the following notions of pattern come together: "use pattern" in language contact (Heine and Kuteva 2005), "linguistic and auxiliary pattern" in linguistic complexity (Dahl 2004), and "morpho-syntactic template" in linguistic typology (De Groot 1990, 2005b).

Abbreviations

A	adjective
ACC	accusative
ASP	aspect
ATTR	attributive marker
CAUS	causative
COP	copula
DEF	definite conjugation
DER	derivational affix
INDEF	indefinite conjugation
INES	inessive
INF	infinitive
INSTR	instrument
MOD	modality
N	noun
NDER	nominal derivational affix
PART	participle
PAST	past tense
PL	plural
POSS	possessive
PRES	present
REFL	reflexive
REL	relative

SG	singular
SUP	superessive
SX	suffix
VDER	verbal derivational affix
V	verb

References

Aboh, E.O. & Ansaldo U. 2007. The role of typology in language creation: A descriptive take. In *Deconstructing Creole*, U. Ansaldo, S. Mathews, & L. Lim (eds), 39–66 Amsterdam: Benjamins.

Bybee, J., Perkins, R. & Pagliuca, W. 1994. *The Evolution of Grammar. Tense, Aspect, and Modality in the Languages of the World*. Chicago and London: Chicago University Press.

Dahl, Ö. 2004. *The Growth and Maintenance of Linguistic Complexity*. Amsterdam: John Benjamins.

De Groot, C. 1989. *Predicate Structure in a Functional Grammar of Hungarian*. Dordrecht: Foris.

De Groot, C. 1990. Morphology and the typology of expression rules. In *Working with Functional Grammar: Descriptive and Computational Applications*, M. Hannay & E. Vester (eds), 187–201. Dordrecht: Foris.

De Groot, C. 1995. Aspect, mood, and tense in Functional Grammar. In *Temporal Reference: Aspect and Actionality. Vol. 2*: *Typological Perspectives*, P.M. Bertinetto, V. Bianchi, Ö. Dahl & M. Squartini (eds), 29–45. Turin: Rosenberg & Sellier.

De Groot, C. 2005a. The grammars of Hungarian outside Hungary from a linguistic-typological perspective. In A. Fenyvesi (ed.), 351–370.

De Groot, C. 2005b. Morphosyntactic templates. In *Morphosyntactic Expression in Functional Grammar*, C. de Groot & K. Hengeveld (eds), 135–161. Berlin: Mouton de Gruyter.

Dik, S.C. 1997. *The Theory of Functional Grammar, Part 2. Complex and Derived Constructions.* Berlin and New York: Mouton de Gruyter.

Fenyvesi, A. 2005. Hungarian in the United States. In A. Fenyvesi (ed.), 267–318.

Fenyvesi, A. (ed.) 2005. *Hungarian Language Contact outside Hungary: Studies on Hungarian as a Minority Language*. Amsterdam and New York: John Benjamins.

Foley, W.A. & van Valin, R. 1984. *Functional Syntax and Universal Grammar*. Cambridge: Cambridge University Press.

Heine, B. & Kuteva, T. 2005. *Language Contact and Grammatical Change*. Cambridge: Cambridge University Press.

Hengeveld, K. 1992. *Non-verbal Predication. Theory, Typology, Diachrony*. Berlin and New York: Mouton de Gruyter.

Hock, H.H. & Joseph, B.D. 1996. *Language History, Language Change, and Language Relationship*. Berlin and New York: Mouton de Gruyter.

Lefebvre, C. 2001. What you see is not always what you get: Apparent simplicity and hidden complexity in creole languages. *Linguistic Typology* 5: 186–213.

Nichols, J. 1992. *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press.

Olbertz, H. 1998. *Verbal Periphrases in a Functional Grammar of Spanish*. Berlin and New York: Mouton de Gruyter.

Rijkhoff, J. 2002. *The Noun Phrase. A Typological Study of its Form and Structure*. Oxford: Oxford University Press

Language complexity and interlinguistic difficulty

Eva Lindström
Stockholm University

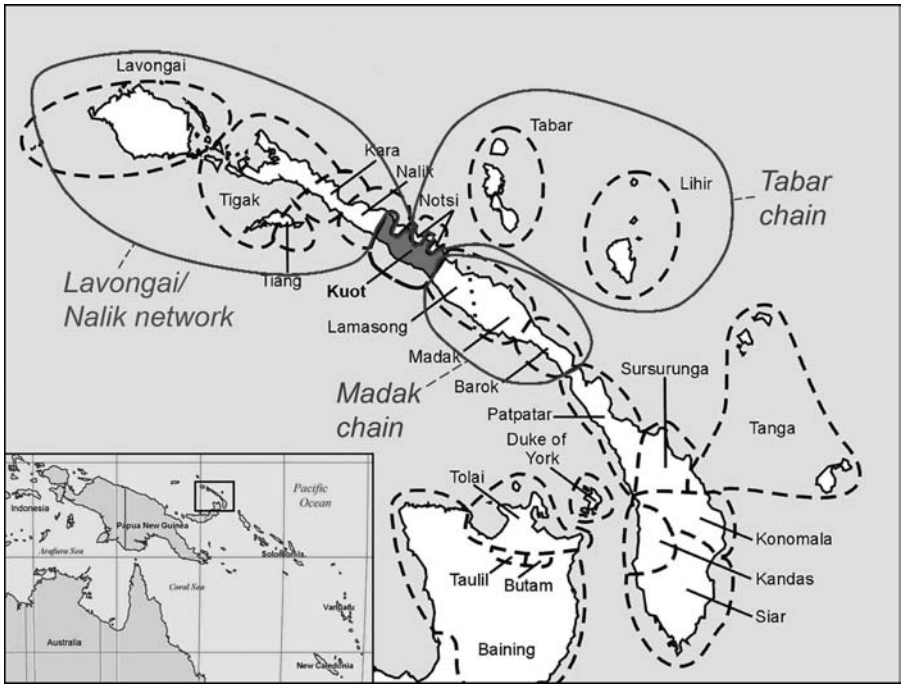
This paper explores the related but distinct issues of linguistic complexity and difficulty, as from the viewpoint of an adult learner. Language complexity is seen as an objective property of a system, which could in principle be computed mathematically, while difficulty is grounded in the particular person who experiences the difficulty, involving factors such as the linguistic categories present and the nature of their marking in the learner's own language. This reasoning will be illustrated with one non-Austronesian language, Kuot, and its three Austronesian neighbours, Nalik, Notsi and Madak, of north-central New Ireland, Papua New Guinea.

1. Background

Kuot is spoken in the linguistically most diverse region on earth: the country of Papua New Guinea has 5 million inhabitants and around 800 languages. In the islands and on the coast of the main island of New Guinea there are over 100 languages belonging to the Oceanic branch of the Austronesian family, a relative new-comer to the region at about 3,500 years ago. Remaining languages are subsumed under the label "Papuan", which, however, is a negatively defined category, implying only that these languages are non-Austronesian and spoken in this region. They dominate in the interior of New Guinea, and around 25 such languages are scattered across northern Island Melanesia. They are assumed, in the main, to derive from languages spoken by the first settlers of the region who arrived more than 40,000 years ago (and presumably in some subsequent waves; Spriggs 1997).[1]

1. I wish to thank one anonymous referee, the editors and Wouter Kusters for feedback which has much improved this paper. I also want to thank Lee Erickson and Bob Lee for letting me use their unpublished data on Notsi and Madak respectively; the latter also for discussion on the language situation. Gunnar Eriksson has read draft versions of this paper and provided useful discussion. Initial parts of this work, as part of the European Science Foundation EUROCORES Programme OMLL, was supported by funds from Vetenskapsrådet and the EC Sixth Framework Programme under Contract no. ERAS-CT-2003-980409.

Kuot (Lindström 2002) is the only Papuan language of New Ireland in the Bismarck Archipelago, indicated by shading in Map 1. It borders on three Oceanic languages, Nalik (Volker 1998), Notsi (Nochi, Lesu; Erickson and Erickson 1991), and Madak (Malom variety; Lee 1978, 1989, n.d.), each from a different subgroup of Oceanic.[2]



Map 1. New Ireland languages

For the last approximately 100 years, there has been a growing lingua franca in the area, the English-lexified creole Tok Pisin (similar to what is sometimes called Melanesian Pidgin or Neo-Melanesian). Essentially every member of the Kuot community is fluent in it, and many children speak no other language. However, children in neighbouring Oceanic-speaking communities typically do speak the traditional languages of their communities, and in fact Kuot children who play with Notsi children will know Tok Pisin and Notsi, but not Kuot. Yet the neighbouring communities are

2. The reference of the term Madak can be confusing, as it is the name of a small language family "the Madak languages", as well as one of its members, the languages being Lavatbura-Lamusong, and Madak. Even within each of these, there is much dialectal variation, and Lee (1978, 1989) cautiously restricts the scope of his observations to the Malom variety, which, unfortunately for our purposes, is non-adjacent to Kuot.

well-nigh identically configured in terms of population size, culture, history, and so forth – if anything Notsi is the smallest of them all, spoken in only two villages.

My impressions of the language ecology of this area derive from a total of around two years of linguistic fieldwork in the period 1997–2004, spent mainly in the Kuot village Bimun on the south-east coast of New Ireland. In that community, Tok Pisin is used in village meetings and church services, and in many everyday conversations, even in households where both adults are full Kuot speakers. Generally, people aged 35 and older are full speakers, 20–25-year-olds are semispeakers, and children are not acquiring Kuot at all (although there is an exception in a closeby hamlet where Kuot is consistently spoken; however, the children switch to Tok Pisin as their main language of interaction as soon as they are old enough to go and play in Bimun). There are often non-Kuots present; some are New Guineans who worked at the former plantation nearby and stayed on, others have married in from neighbouring language areas or further away, and some are temporary visitors. Politeness therefore often requires the use of Tok Pisin or another language shared by the persons present.

Speakers of Nalik, Notsi, and Madak languages unanimously declare that Kuot is inordinately difficult, while their own languages are "easy". Currently, few people who move into the Kuot-speaking area ever acquire an active competence in Kuot, although most learn to understand it. Conversely, Kuot speakers often acquire quite a good competence in neighbouring languages if they are in contact with the speakers. This appears to have been the situation historically too, although more non-Kuots knew some Kuot then; today, cross-linguistic communication does not require bilingualism other than in Tok Pisin.

Another factor behind the notion that Kuot is difficult is most certainly that the other three languages considered are related to each other (and to all other languages on the island, except Kuot). Even though they belong in different subgroups of Oceanic, they share a relatively recent common ancestor, and much lexical and grammatical material will be familiar from one language to the other. Nalik and Notsi are closer to each other than either is to Madak, but the distance between any two of them is not greater than that between English and German. To these, Kuot may be said to be like Finnish or Hungarian: of a very different origin and with a very different structure to its grammar and morphology, but culturally and historically very much integrated with them, and having a large body of lexical loans. There are of course salient differences between the New Ireland situation and the European case, such as the fact that the New Ireland languages are small in numbers of speakers, and spoken in even smaller settlements of swidden horticulturalists in societies with very little social stratification. In spite of indications that personal mobility was rather limited traditionally, it seems likely that a larger proportion of speakers of each language would have been in contact with speakers of other languages than in Europe. The assumption of extended contact across language boundaries is supported in oral history; in the fact that clan territories straddle language boundaries; in the fact that a fair amount of kin terminology is shared between unrelated languages; and in the fact that there is what

Ross (1994) has called a phonological alliance in this part of New Ireland, with several shared phonological rules and restrictions (Notsi participates only to a limited degree in this Sprachbund).

The *Ethnologue* (Gordon 2005) gives the population for each language but the information is from different years and different sources and therefore hard to compare: Kuot, population 2400 (2002); Nalik, 5138 (1990); Notsi, 1836 (2000); Madak (the language, not the family; cf. note 2), 3000 (1985). Note, however, that language boundaries constitute census boundaries and that these figures should be taken to indicate the number of residents within the census division defined by each language, rather than the actual number of speakers.[3]

It deserves pointing out that the Oceanic languages in question are by no means simple from a linguistic point of view. And, of course, the fact that the Oceanic languages in central New Ireland are more similar to each other than any of them is to Kuot does not in itself constitute a reason why it should be more difficult for, say, a Notsi speaker to acquire Kuot than for a Kuot speaker to acquire Notsi. Yet the situation is not symmetrical.

2. Complexity and difficulty

So how could we investigate the claim that Kuot is more "difficult" than its neighbours?

The concept of *complexity* that I am operating with is structural, viewing language, in this case grammar, as a system whose complexity could in principle be computed mathematically. Although linguists may disagree on how to formulate the metric, it is still the case that, for example, irregularity in paradigms is inherently more complex mathematically than regularity, in the sense that more rules are required to describe it, or more machinery of some kind, regardless of the descriptive framework.

Thurston (1987) does not make an explicit distinction between system-based and user-based perspectives, but writes in the context of learnability, and he expresses the concept of complexity quite well in the following passage:

> By way of terminological definition, the simplest language would have a perfect one-to-one correspondence between a unit of meaning and the form encoding it. It would have no stylistic or sociolectic variation. It would have the fewest possible number of grammatically marked obligatory distinctions and a small morpheme inventory from which the full lexical inventory could be predictably

3. The claim in the *Ethnologue* that Kuot is "vigorous on the west coast" is not supported by my own observations. In general, it would have been desirable to have more detailed sociolinguistic information on the Oceanic languages discussed here, but unfortunately such information is not available to me at the time of writing.

> derived by a small set of regular derivational rules. Taking full advantage of analogy, each rule would be applicable wherever it made logical sense; and there would be one rule per function. The most complex language, on the other hand, would have all the things we use to torture students of introductory linguistics – allophony, allomorphy, unpronounceable consonant clusters, gratuitous morphophonemices, and unpredictable suppletion. It would have an enormous morpheme inventory with many near-synonyms differing in slight shades of meaning with implications of degrees of formality and socioeconomic status. The most complex language would also have a lexicon which relied heavily on a large number of opaque idioms. (Thurston 1987: 41)

(Note, however, that in this paper we will concentrate on morpho-syntactics and touch only briefly upon vocabulary and social factors.)

Difficulty, on the other hand, is related to complexity but differs from it in that while complexity is a property of a system, difficulty is subjective. That is, something is difficult or not difficult for a particular person. Difficulty therefore depends on the individual we take as our starting point. If I am Swedish and learning Estonian, it is very difficult as the two languages are very different; if I am Finnish it is a whole lot easier as many words and structures are closely related, quite independently of the complexity of the systems involved.

In principle, categories that are present in one's own language, or other languages one knows well, should be easier to acquire in a new language, such as definiteness and number marking (disregarding for the moment the form of the marking). However, this is not always the case, as for instance with gender, the assignment of which differs quite a bit even between fairly closely related languages.

In this paper, our imaginary learner is an adult speaker of one of Kuot's Oceanic neighbour languages, who wishes to acquire Kuot in order to use it in communication.

The definition of difficulty here is dependent on the adult acquisition perspective, and other properties of language, such as aspects of language use, would presumably require a different analysis. For example, irregular paradigms could be beneficial in language input processing for individuals who already have a high competence in the language, as some cases of irregularity can most likely help avoid ambiguity, shorten search times for a match in the mental lexicon of a listener, etc., depending on your favourite model of language processing.

Other authors have made very similar distinctions to that which I make between complexity and difficulty. Dahl speaks of system complexity (2004: 42–43) which is independent of use (p 39), and difficulty (pp 39–40) which is anchored in an agent. Miestamo (2006; this volume) distinguishes absolute complexity as a property of a language, and relative complexity which is defined with regard to a language user. I agree with the distinctions made by these authors.

It may be noted that Kusters (e.g., this volume) differs somewhat, making the concept of complexity agent-related, although it should be noted that his agent, also an

adult L2 learner, is depersonalised by the assumption of no prior linguistic or cultural profile.

3. "Counting difficulty"

In this paper I have in mind a simple left-to-right procedure, imagining the path through a clause from the beginning to end, checking how many choices are encountered at each stage, roughly according to the following list:

- how many categories are morphologically marked (obligatorily/typically)?
- how large a paradigm is there per category?
- how much variation (esp. irregularity) is there in each paradigm?
- how many of the grammatical categories are present in the learner's language?
- how similarly are they marked in the learner's language?

The first three of these relate to complexity, the last two to difficulty.

Even such a simple left-to-right method is not unproblematic. For example, if an earlier choice limits the selection of possible forms at a later stage in the clause, how should we count the number of choices at the second stage?[4]

And what about categories that are not really compulsory but without which the construction is unidiomatic? We will return to such issues below.

Here, I will not perform a full count, but illustrate the sorts of structures and paradigms we would have to take into account if we were to attempt this kind of analysis, and discuss the sorts of problems that arise.

A few caveats are in order. I have very limited experience of the languages discussed except for Kuot, and have relied on other sources. There is a published grammar only for Nalik (Volker 1998). For Madak and Notsi there are unpublished typescripts which the authors, Robert Lee and Lee and Laurinda Erickson, respectively, have kindly let me use. I have not had access to text in Nalik.

A hazard in all types of typological comparison is of course the fact that the number and types of categories postulated are analysis-dependent, and some of the summaries to follow represent my own interpretation of the materials available to me.

Table 1 gives a short overview of some grammatical categories and the numbers of forms marking them in the four languages we are considering, plus the creole Tok Pisin. Parentheses for plural marking on nouns indicate that there is a form to mark non-singular but it is a separate word. The plus sign (+) shows that there are forms for the category but that counting is not applicable; the '3+' for Kuot verb classes is because there are three intransitive classes paired with four transitive classes.

4. Cf. Dahl's concept of choice structure (2004: 46–50).

Table 1. Some categories and numbers of forms

	Kuot	Nalik	Notsi	Madak	Tok P.
gender	+	–	–	–	–
NSG marked on N	13	(1)	(1)	4	(1)
S marked on V	30	8	15	11	–
O marked on V	19	–	–	–	–
V classes	3+	–	–	–	–
tenses	2	2	3	5	2
inal. poss	12	9	4	3	–
alien. poss NP–NP	44	1	15	1	1

Some of these categories will be presented in more detail below. First, however, we will have an example of a transitive clause with two full noun phrases from each language (Tok Pisin excluded). Noun phrases and verb phrases are given in brackets.

(1) Nalik
[*A mun finau*] [*di wut buak*] [*a vaal*]
ART NSG thief 3NSG come break ART house
'The thieves came (and) broke (into) the house.'

(2) Notsi
[*Yia-ka tamat*] [*a-choka*] [*kariin ko*]
DEM-INDEF man 3SG.SUBJ-spear huge fish
'One man speared a huge fish.'

(3) Madak
[*la-va-kin*] [*di-t-kis gugu*] [*leng-kaxi atdi*]
NM-PL-woman 3PL-CN-sit make NM.PL-basket their
'The women are sitting making their baskets.'

(4) Kuot
[*o-ikat-oŋ*] [*Adam*] [*muabari aŋ*]
3F.OBJ-check-3M.SUBJ.NFUT Adam sun/watch(f) 3M.APOSS.3SG
'Adam checked his watch.'

The Oceanic languages all have SVO word order, while Kuot has VSO. Kuot further cross-references objects of transitive verbs, while the others do not.

We will now turn to examine some subsystems of the grammars of these languages in a little more detail.

3.1 The verb phrase

This section investigates the variability in marking within the verb phrase. Appendix I gives more detailed data for each language. The Oceanic languages are similar to one another in that all have a sort of bracket system, where the verb phrase is introduced by

an obligatory subject marker, followed by five to six slots for optional markers before the verb stem (for Madak it is made explicit that only three markers at a time are possible). Preverbal markers include tense and aspect, mood, adverbial markers, and negation. After the verb stem more optional material can be added, more or less bound to the verb; primarily serialising verbs, and Nalik and Notsi have transitivizing suffixes. The languages differ from each other in the number of slots, the number of categories in each slot and to some extent the order of items, and in the forms that mark each category.

Kuot does not have the bracket system, as the subject marker is more closely integrated with the stem, and further occurs in different positions relative to the stem depending on the verb class, and object markers also occur in different positions depending on verb class and grammatical person; see Appendix III. Other markers indicate categories very similar to those of the neighbouring Oceanic languages but with some peculiarities. First, a number of enclitics may attach to the first constituent of a phrase; second, the continuous marker *mən*, as well as full adverbs, may go in various positions in the verb phrase.

Kuot and Nalik differentiate two tenses, non-future and future, while for the Notsi and Madak it is not clear to what extent the use of the markers indicated for their rather more elaborate tense distinctions are strictly required or merely possible.

In terms of complexity, the main difference between the languages is that many more post-verbal slots are reported for Nalik and Notsi than Madak or Kuot. In terms of difficulty, the verb phrase structure of the Austronesian languages would transfer quite easily from one language to another, although forms and details would of course have to be learned. Kuot, on the other hand, does not have a very complex verb phrase structure but the variable position of some markers will be unfamiliar, and the many subject marking strategies together with the fact of object cross-referencing a definite challenge to a speaker of an Austronesian language.

3.2 Pronominal systems

All the languages under consideration have inclusive/exclusive forms of non-singular pronouns, at least some degree of expression of dual number and sometimes trial or paucal. All have some degree of pronominal marking on prepositions. Differences include the presence or absence of a separate set of subject markers parallel with the independent pronouns, and how many grammatical numbers are used. Full paradigms are given in Appendix II.

Nalik has the smallest number of pronominal forms among the languages studied here. As is common among Oceanic languages, subject markers are different from independent pronouns; the latter are also used in object functions. The only pronominal forms that interact with prepositions are those for alienable possession. Although dual and paucal numbers exist, they are used only rarely, and are based on non-singular (sometimes alternatively singular) forms. Inalienable possessive constructions have all but disappeared from use.

Notsi pronominal forms make up a large system, with some consistency across categories but not full regularity. In Notsi, independent pronouns are used as subject markers, and may function as the only exponent of the subject argument of a verb. There are some special forms used with prepositions which are also used as independent object pronouns, overlapping to a large extent with the subject/independent forms. Special forms are used for the benefactive, mostly derived from other forms except in the first person, and there is also a set of related benefactive forms used when food is the gift. Dual and trial forms contain components of the numerals *lua* 'two' and *tuul* 'three'.

Madak has a fairly large set of pronominal forms, differentiating independent forms for subject and object, and having a separate set for subject marking on the verb. Again, special forms go with particular prepositions, most of which are derived from other forms in the system. There are three numbers: singular, dual and plural.

The Kuot pronominal system has many forms, due in part to the fact that verbs cross-reference both subjects and objects. The forms used differ between the verb classes, and a class of adjectives take a partly different set of markers again (cf. also Appendix III). Prepositions are indexed with pronominal forms if there is no nominal in the prepositional phrase (and sometimes even if there is). Markers for inalienable possession are closely related to object prefixes, which are also nearly identical with preposition indexers. There is a large system of forms marking alienable possession, indexing all 12 pronominal categories of possessors, as well as possessees (with some syncretism of masculine and feminine possessees); forms for singular feminine possessees are similar to subject forms of verb class I. There are three numbers: singular dual and plural, and two genders: feminine and masculine, distinguished only in the third person singular.

It is hard to say which is the most complex of these pronominal systems. Nalik and Madak have relatively small systems; the Notsi system makes more distinctions and has four numbers but recycles many of the forms adding suffixes to form new categories, and does not distinguish subject forms from independent forms. Kuot expresses a large number of categories but recycles a fair amount of material between them, such that it is often possible to recognize the person and number of a form new to the listener. In terms of difficulty, Kuot causes speakers of neighbouring languages a lot of trouble by distinguishing feminine and masculine in the third person singular, by marking objects on verbs, and by using different positions of marking and somewhat different markers in different verb classes. Of these complications, it appears that gender is the hardest to master; even very fluent learners who handle verb classes and object agreement without thinking make frequent gender mistakes.

3.3 Nominal morphology and noun phrases

Notsi and Nalik have no nominal morphology. Number marking is done with a free morpheme.

Madak, on the other hand, has quite a large set of prefixes for nouns. They are prefixed articles (or noun markers), which can be combined with morphemes that mark plural and a number of other categories such as diminutive and parthood. The assignment of forms to particular nouns is not fully transparent, and noun classes (declensions) can be postulated among common nouns on the basis of prefix selection; if based on the plural prefix there is one big class and four smaller ones.

Kuot nouns fall into 13 declensions based on plural formation. For about half the nouns, the plural is formed by suffixing *-(V)p* to the noun stem. Most of the rest end in particular forms which are removed before the plural ending is added (e.g., *kiraima* (sg) – *kiraip* (pl) 'nail, claw'; *kikinəm* (sg) – *kikip* (pl) 'ear'). The dual *-ien* is added onto the plural (except in many dyadic kin terms which have special dual forms).

Unpredictability and irregularity as in Madak and Kuot automatically give complexity as set out above. For speakers of Nalik and Notsi marking plural on nouns at all is an unfamiliar thing to do and they would have to acquire it as a linguistic habit so to speak. Speakers of Kuot and Madak expect plural marking but the irregularity would still cause problems.

Aside from morphology, the syntax of noun phrases appears to be roughly equivalent in complexity across all four languages.

4. Discussion

The above inspection of three grammatical subsystems in four central New Ireland languages points to one of the problems with this type of exercise: the different areas do not necessarily lead to the same ranking of languages. For instance, Kuot and Madak verb phrases are not very complex while nominal morphology is, and the opposite holds for Nalik and Notsi.

Further, it is not always clear which criteria should be emphasised; for instance the number of functions encoded in the pronoun system vs. consistency of form across each person–number category.

The left-to-right procedure in this crude form also fails to capture interdependencies between items in a clause, such as agreement marking on demonstratives and other constituents. Nor is it clear how to treat the variable affixation order that comes with the different verb classes in Kuot.

Regardless of how we determine complexity, however, it is clearly the case that much more than complexity figures into difficulty, and we will have one more demonstration. Let us imagine a Notsi speaker who is going to approach the production of a Nalik clause. Let us say he wanted to produce the clause given in (1) above:[5]

5. For 'he' in this scenario, read 'he or she'.

(1) Nalik
[*A mun finau*] [*di wut buak*] [*a vaal*]
ART NSG thief 3NSG come break ART house
'The thieves came (and) broke (into) the house.'

We will assume that the lexemes are known. The word order is SVO as in Notsi so that is not a problem. Almost all Nalik nouns take a non-specific article (noun marker) *a* (or *na*) in most contexts, while Notsi has no articles, so our speaker may or may not remember to put that in. In the first noun phrase he would probably not have a problem with the plural marker *mun*, as Notsi has *no* in the same place and function.

In the verb phrase, Nalik subject markers are different from the independent pronouns, which may cause trouble. Now this is the sort of area that requires careful thinking if we wanted to calculate difficulty. Do we imagine our Notsi speaker confusing subject markers only with independent subject pronouns, or with third person plural forms from across the whole pronominal paradigm? As for serial verbs, there are similar constructions in Notsi, so again no problem. The last noun phrase has only an article and a noun so only the article would make the construction different from a Notsi phrase.

Now let us imagine our Notsi speaker attempting a Kuot clause, the one given above as (4):

(4) Kuot
[*o-ikat-oŋ*] [*Adam*] [*muabari aŋ*]
3F.OBJ-check-3M.SUBJ.NFUT Adam sun/watch(f) 3M.APOSS.3SG
'Adam checked his watch.'

Here the word order is VSO, which will be unfamiliar, although it is possible to front an argument of the verb (if so the construction is marked by a form *lə* after the fronted constituent: *Adam lə o-ikat-oŋ muabari aŋ*), so our Notsi speaker could have a derived SVO structure if he wanted. Whether he does or not, the verb constitutes our first real problem. For one thing, the verb stem selected decides what affixation order is possible, and what affixes are to be used (cf. Appendix III). In this case, the verb is of class I, so the speaker has to know the forms for that (again, it is unclear whether all other forms for the same pronominal category should be regarded as a source of confusion, or just those for subjects, or those for subjects as well as objects, given that both types are verbal affixes). The singular masculine subject marker further comes in two versions, future and non-future. In addition, object marking has to be considered, obligatory in a transitive construction, and a novel thing for a Notsi speaker, who would only be familiar with rather limited transitivity marking. In this case, we have a singular object, meaning that the distinction between feminine and masculine is relevant, also an unfamiliar concern from a Notsi point of view.

In the object noun phrase there is a noun and an alienable possessive pronoun; the possessive is an unfair complication for our comparison as the Nalik phrase did not involve possession, but we may comment that the possessive pronoun contains

the categories for singular masculine possessor and singular possessee (the singular masculine possessor form happens not to index the gender of the possessee), so this is quite a difficult area of Kuot grammar to master for a learner.

To sum up, the Nalik sentence requires an article (choice of four, if Ø, *na* and a specific article *ta* are included), plural (choice of one or none), a subject marker (choice depends on assumptions of what are confusable items), and another article for the last noun phrase.

To produce a correct Kuot sentence, the gender of the object noun has to be known and cross-referenced, the affix form and place depending on the class of the verb, which also governs the form and position of the subject cross-referencer, some of which also pay attention to tense (range of choice again depending on our assumptions). Even for producing an incorrect Kuot sentence, the differing orders of obligatory verbal affixes constitute a hurdle.

4.1 Evasion

So far so good. But we have been cheating a little by considering a ready-made sentence to match. A question that arises is how we are to think of all the optional marking for tense and aspect in the verb phrase, for instance; they add to the possible choices on a systemic level but would not always be "active" or relevant in a normal speech situation. Non-obligatory items could perhaps be evaded in early stages of acquisition, and in fact I have understood that Kuot speakers conversant in Lavatbura-Lamusong (the nearest language in the Madak group which they picked up from school mates in the mixed school) get by with a reduced system of noun markers.[6] The possibility of evasion as a mechanism in early production should therefore also be considered, but is very difficult to evaluate reliably, and attitudes among the speakers of the target language are likely to be relevant too. That is, if a learner uses no tense or aspect marking in a Nalik verb phrase even when it is expected, will it be accepted by Nalik speakers as communicatively valid? Will Notsi speakers accept learners perhaps using the plural where the trial, which is not in the other languages, had been expected? The case of Kuots reducing Lavatbura-Lamusong noun markers suggests that this is quite acceptable learner behaviour. The fact that Kuot pronominal marking on verbs, especially in verb classes II and III, is much more tightly integrated with the stem than in the Oceanic languages means that there is no way of avoiding cross-referencing, and the fact that this is done in several different ways in turn means that you can say very little without quite a bit of knowledge of this area of grammar. It would seem that even

6. I do not speak or understand Lavatbura-Lamusong but have heard particular speakers of Kuot speaking it seemingly competently. However, discussing it later, they were unaware of meaning differences beyond grammatical number between nouns prefixed in different ways (I knew of these from interviewing Lavatbura-Lamusong speakers).

speaking *bad* Kuot requires more extensive knowledge of the system than speaking bad, or just limited, Nalik.

4.2 Mismatches: Category content and cooccurrence restrictions

Although a learner's language may have particular categories, and even code them in a similar way to the target language, there can be mismatches in application. One such item in these languages concerns the use of an inalienable possession construction. In Madak (which has inalienable forms only for singular possessors) and Notsi, it is used for some kin terms and some body parts. In Kuot, inalienable possession is used for all types of parthood, but alienable possession marking is used for kin (and in Nalik, the inalienable possessive construction has essentially fallen out of use altogether).

Another area of mismatch is the cooccurrence restrictions on markers of different grammatical categories. For example, in all the languages, the marker for future time reference has irrealis connotations, but the use of the future marker with negation operates differently, such that in Kuot there is a disassociation of future marking and negation marking, while in Nalik, the future/irrealis marker *pe(n)* is obligatory with negation at least for younger speakers (who have reinterpreted it as a pure irrealis marker; Volker 1998: 59), while another future marker *na* can be used freely with negation. In Nalik, the durative marker may not be used with negation, while in Notsi it may, and so forth. It is not clear to me at this point how this type of mismatch would count in our difficulty scheme.

4.3 Form transparency

As mentioned above, the languages are part of a phonological alliance, sharing processes such as the lenition of voiceless stops in intervocalic position. Not all the processes and restrictions are identical in all the languages but many recur. Phonology, therefore, is not a great obstacle to interlinguistic understanding in these languages.

But Kuot and Madak have some additional processes. Madak has processes of vowel harmony that occur between stems and markers, and between concatenated markers, as well as vowel deletion in stems depending on syllabification with prefixes (e.g., *la-ban-tixin* NM-thin-woman 'old woman', but *la-tkin* NM-woman 'woman'; *la-pke* 'mud', but *la-xam-pixe* 'much mud').[7] However, these appear to be more of a hindrance to impeccable production than to comprehension, as speakers appear to be aware of the shape of the underlying form in most cases, and have no trouble giving the unaffixed form of a noun if asked, and even Kuot speakers near the language boundary tend to know quite a bit of vocabulary from the Madak language closest to them (Lavatbura-Lamusong), and have a clear idea of the basic forms of nouns.

---

7. In Madak orthography, ⟨x⟩ stands for the voiced velar fricative [ɣ], which is an intervocalic allophone of /k/; as mentioned above this is a process common to several of these languages (cf. Ross 1994; Lindström in prep.)

In Kuot, there are morpho-phonological processes that take place between stem-initial vowels and vowels of the subject prefixes in verb classes II and III. These often obscure the forms of the prefix and/or the verb stem which may make it more difficult for a learner to make sense of the system, for example /i-onəma/ is pronounced [jonəma] 'she sits/lives', but /u-onəma/ 'he sits/lives' becomes [unəma], and /pa-onəma/ 'we (excl.) sit/live' becomes [ponəma].

Processes such as those in Madak and Kuot conspire to obscure the form of the stem and sometimes the affixes, which is of course unhelpful to a learner.

For Kuot, the presence of morphological processes like the subtraction of many noun singular endings in plural the formation (see above), and the fact of variable affixation order for verbs is not going to help a learner: the observable representation of morphemes is not constant. From the point of view of an Oceanic speaker, the variation associated with gender may well seem random too.

4.4 Vocabulary

Another area of potentially variable complexity is vocabulary, which should be mentioned even if not computable within the simple metric explored here.

There is far-reaching alignment in the types of meanings encoded and the scope of meaning of many words across the languages. As an example, the word normally translated as 'kill' is typically not result-oriented but action-oriented, implying severe beating up and such, so that an utterance like 'he killed her so she was really upset' is perfectly logical. This is reflected in Tok Pisin too, the structure of which is of course based on the local languages in this region. Another item similarly reflected in Tok Pisin is a word *inap* (from English 'enough') which means 'enough' as well as 'to align' (in the transitive form *inapim*), 'to reach', and 'to be able to', and the same polysemy is found in Kuot *puo* and *pupuo*; Notsi *pupua* means 'enough' and 'be able to'; Nalik *faraxas* is 'be same as' and 'be able to'; and Madak *epovo* means 'be enough' and 'be able to' (for each Oceanic language, these are the meanings I have been able to find; there may be more). There are many many similar cases, pointing to the extended contact between speakers of the different languages over time. However, it is possible that Kuot has a richer vocabulary in the sense of having more lexical stems: speakers of Kuot and speakers of the neighbouring Madak language Lavatbura-Lamusong agree that Kuot has more words and "little meanings". For example, while Lavatbura-Lamusong frequently derives words for fruits, tree, leaves etc. from a single lexical item, Kuot often has separate lexemes for each, and there are indications that there is more resolution also among verbs. The reason why Kuot should have more words than its neighbours is that it appears to have borrowed from several of its Oceanic neighbours and given the acquired words different functions. This information on vocabulary size is somewhat anecdotal and would need to be researched in a principled way before being admitted as support for a claim that Kuot is more difficult to acquire than its neighbours.

The influence of many languages is generally accepted as the reason for the large size of the vocabulary of English. However, in the English case there has been a levelling (simplification) of the grammar which is often attributed to the presence of second-language speakers in leading positions. It is interesting to note that such levelling appears not to have taken place in Kuot to the same extent, and perhaps the reason is that fewer outsiders, or too small a number of outsiders, have learnt it.[8]

4.5 Popular judgment of difficulty

Kuot is not the only Papuan language in a setting of Austronesian languages to be judged difficult by its neighbours. At least two other cases are discussed in the literature: Anêm of New Britain and Waskia of Karkar Island, Madang Province.

Thurston (1982, 1987, 1992) writes about Anêm, a Papuan language of West New Britain which is surrounded by Oceanic languages, that "[a]ll [Austronesian]-speaking peoples in the area consider Anêm far too difficult to learn", and he quotes Lusi speakers saying that the "Anêm language is heavy in our throats" (1982: 11). Thurston (1987), much like the above presentation of Kuot and its neighbours, gives paradigms of subject and possessive markers, and discusses stem form transparency, phonological processes, the number of grammatical distinctions made, and other factors, and convincingly shows that Anêm is more complex than its neighbours, and that this is one important reason why it has few learners. Other factors include some of the Oceanic languages being lingua francas with little identificational value for the original speakers, while Anêm has become a marker of its speakers' identity and traditional culture. Yet, on the levels of syntax and semantics, all the languages of north-west New Britain have converged so that their syntactic and semantic structures are virtually identical, as a result of long association with intermarriage, shared culture, and multilingualism as well as language shift by the original inhabitants to the languages of the Austronesian newcomers.

It may be noted that the situation of Kuot differs in that it has no emblematic function, and although semantic structures are very similar to those of neighbouring languages and most of the grammatically encoded categories are the same, it is morphologically and syntactically distinct, and it has even retained VSO constituent order in the face of very dominant SVO among its neighbours.

---

8. An area of grammar that may be seen as a possible exception is the verb classes and adjectives (cf. Appendix III): the only open verb class is that of the least morphological complication, likely derived from a noun+light verb construction. However, there are several hundred stems still in use of classes II, III and adjectives, including many basic verbs, auxiliaries and serializing verbs. The loss of productivity of these classes could tentatively be accounted for by the presence of non-native speakers, and it may be noted that today's semi-speaker young Kuots typically have trouble with less common verbs of those classes.

McSwain (1977) gives an anthropological account of the population on Karkar Island, which is divided between speakers of Papuan Waskia and Austronesian Takia. She notes that the traditional social structure of the two groups was identical, but that there was a clear social division based on the linguistic difference. She quotes speakers talking about the difference: "'Takia', they said, 'lies easy in the mouth and can be learned quickly; Waskia lies heavy and is a difficult language to master'". These perceptions are even encoded in a myth about two brother deities from whom the island's population descended, and from whom the Takia inherited their 'melodious' language and the Waskia their 'cumbersome' language (McSwain 1977: 4). Interestingly, these very languages are famous for being morpho-syntactic carbon copies of one another – what Ross (e.g., 2001) has termed metatypy: the total convergence of structure but with distinct lexical and morphological forms. In this case, the Papuan language, Waskia, was the primary inter-community language and it is Waskia's structure that has been copied by Takia. According to Ross (pers. comm.), Takia is phonologically more complex, and a case for Waskia being morphologically more complex could be made but is not very strong – yet Waskia is considered more difficult.

In the examples just shown, it seems that the perception of Anêm as difficult could be put down to complexity. But for Waskia, it is hard to argue that it is more complex than Takia, yet the perception persists. In the case of Kuot, we could well imagine that Kuot would be considered difficult on the basis of having gender alone, even if it were equivalent in complexity or even simpler than the surrounding languages in all other respects. It is clear that there is reason to be cautious regarding popular judgments of difficulty, and that these cannot always be reduced to complexity.

4.6 Extra-linguistic factors

Naturally, language structure is not all that goes into any particular learning situation. Social and cultural factors involve items such as the status of the target language, but also attitudes towards imperfect language use by a learner, and expectations on individuals of monolingualism or multilingualism. There would also be socio-cultural differences in learning styles between different societies. On top of these factors, individuals within the same society differ from one another. They may have different levels of motivation for learning a language, for instance depending on business interests and their personal social networks. Individuals will vary in the amount of exposure to the target language they have had, and during what period of their lives, and individuals further have varying talent for and personal interest in language learning.[9]

9. I fear that mentioning individual differences in talent for language acquisition is as politically incorrect as it was until recently to imply that languages may differ in complexity. However, I consider it an empirical fact that people differ in this respect, both in Western society and in the Papua New Guinea society where I have worked.

Each of these factors warrant research and papers and books of their own, and I just mention them here to indicate that a study of difficulty in terms of language structure is never likely to have predictive precision, although it may provide a baseline take on the learning enterprise.

Neither Kuot nor Tok Pisin enjoys any particular social status. It may be added here that the vernaculars of this area appear never to have been important for social identity, and they did not even have names until the 1950s. Imperfect language use appears largely to be tolerated, and multilingualism was common.

5. Conclusions

If we accept the premise that language complexity can be computed in an objective fashion (utopian though it may be at present), it follows that complexity can also be compared between languages and some languages can be found more complex than others, in terms (minimally) of properties such as paradigm size and regularity.

Some types of complexity are likely to equal learner difficulty, perhaps especially paradigmatic irregularity, suppletion and the like, which mean that forms cannot be regularly derived but have to be learnt individually.

Other types of complexity, such as categorial distinctions, figure as difficult primarily when the learner's language does not have the same distinctions, or they are marked very differently. While the different verb affixation orders of Kuot would presumably always be a challenge to a learner, the fact of object cross-referencing is likely to be less of a problem for someone whose own language has object cross-referencing; it would be a matter of remapping a familiar function rather than acquiring an entirely new kind of grammatical behaviour.

As we have seen, e.g., the syntax of Kuot verb phrases is not necessarily more complex than in the neighbouring languages, but there is still reason to assume that a speaker of a neighbouring language would find even this area a lot more difficult to learn than in any of the other neighbouring languages, due to the lack of overlap of marking structures.

Even if we accept that the structures of two languages can be compared in such a way as to reveal wherein the difficulties for a learner from a particular linguistic background will lie, we will still have only a rather crude tool for gauging the difficulty in any actual learning situation, as in any concrete situation a variety of non-negligible extralinguistic factors come into play, including social, cultural and individual variables. With the societies considered being socio-culturally so homogenous, it would be hard to explain the more rapid decline of Kuot compared with its neighbours without reference to the language itself. I hope to have shown that Kuot is more complex in several areas, but, above all, difficult for speakers of neighbouring Oceanic languages.

It is possible that the Kuots have been just that little bit more used to switching into other languages to accommodate non-Kuots, and that this habit is enough to tip the

balance such that the language is about to be abandoned altogether. The Oceanic languages surrounding Kuot are by no means safe from the threat of extinction, but so far children in most places appear to learn them, unlike Kuot children.

Abbreviations

APOSS	alienable possessive
ART	article
CN	continuous aspect
DEM	demonstrative
F	feminine
INDEF	indefinite
M	masculine
NFUT	non-future
NM	noun marker
NSG	non-singular
OBJ	object
PL	plural
SG	singular
SUBJ	subject

References

Dahl, Ö. 2004. *The Growth and Maintenance of Linguistic Complexity*. Amsterdam/Philadelphia: John Benjamins.

Erickson, L. & Erickson, L. 1991. Grammar Essentials of the Nochi Language of New Ireland Province. ms.

Gordon, R.G. Jr. (ed.) 2005. *Ethnologue: Languages of the World*, Fifteenth edition, Dallas, TX: SIL International. Online version: <http://www.ethnologue.com/>

Lee, R. 1978. Madak Grammar Essentials. ms.

Lee, R. 1989. The Madak verb phrase. *Language and Linguistics in Melanesia* 20(1-2): 65–114.

Lee, R. n.d. The Madak Noun Phrase. ms.

Lindström, E. 2002. Topics in the grammar of Kuot – a non-Austronesian language of New Ireland, Papua New Guinea. PhD Dissertation, Stockholm University.

Lindström, E. In preparation. Phonological convergence in New Ireland: Kuot and its Austronesian neighbours.

McSwain, R. 1977. *The Past and Future People – Tradition and Change on a New Guinea Island*. Melbourne: Oxford University Press.

Miestamo, M. 2006. On the feasibility of complexity metrics. In: *Finest Linguistics. Proceedings of the Annual Finnish and Estonian Conference of Linguistics, Tallinn, May 6-7, 2004* [Publications of the Department of Estonian of Tallinn University 8], K. Kerge & M.-M. Sepper (eds), 11–26. Tallinn: TLÜ.

Ross, M. 1994. Areal phonological features in north central New Ireland. In *Language Contact and Change in the Austronesian World* [Trends in Linguistics. Studies and Monographs 77], T. Dutton & D.T. Tryon (eds), 551–572. Berlin: Mouton de Gruyter.

Ross, M. 2001. Contact-induced change in Oceanic languages in North-West Melanesia. In: *Areal Diffusion and Genetic Inheritance: Problems in Comparative Linguistics*, A.Y. Aikhenvald & R.M.W. Dixon (eds), 134–166. OxfordNew York: Oxford University Press.

Spriggs, M. 1997. *The Island Melanesians*. Oxford: Blackwell.

Thurston, W.R. 1982. *A Comparative Study in Anêm and Lusi*. Canberra: Pacific Linguistics.

Thurston, W.R. 1987. *Processes of Change in the Languages of North-Western New Britain*. Canberra: Pacific Linguistics.

Thurston, W.R. 1992. Sociolinguistic typology and other factors effecting change in north-western New Britain, Papua New Guinea. In *Culture Change, Language Change – Case Studies from Melanesia*, T. Dutton (ed.), 123–139. Canberra: Pacific Linguistics.

Volker, C.A. 1998. *The Nalik Language of New Ireland, Papua New Guinea*. New York: Peter Lang.

Appendix I. Verb phrase

The Nalik verb phrase

S	COUNTERFACT	TNS/ASP	REC	DUR/NEG/ LOC	CAUS/ REDUP	V	INC/SER	DL/PC	TRANS	FOC	COMPL
	lek	ANT *tabung* FUT *na* INCEPT *vala* HAB *runa*	*vara=* *vavur*	DUR *i, t* NEG *pe(n)* LOC *su, o*	CAUS *fa*		[inc. O; DO; ser V]	DL: *-(y)a* PC: *-(t)al*	*-ing*	*ang*	*faanong*

The Notsi verb phrase

S	TNS+		ASP+		NEG	MOOD+		REC	V	TRANS	ADV	MOTION	TRANS	LIM	AUG
	FUT	*ba*	PERF	*se*	*kap*	IMPER	*na*	*e-*		*-i*	. . .	*la* 'go'	*ngin*	*mu*	*buka*
	PST	*ti-*	DUR	*mager*		APPR	*soro*				. . .				
	CERT	*ta*	"CONT"	*lo*		PERM	*le*								
			SEQ	*ga*		INDEF	*ka, mara*								
			+			HAB	*tigiri*								
						ABIL	*pupua*								
						INTS	*suuk*								
						INTS	*mun*								

(Causative *a-* occurs after mood but it is not clear in what relation to reciprocal).

The Madak verb phrase

S	IRR	TNS	NEG	ASP	X	CONT	V	MODIF
	gi-	YPST *ta* RPST *ga* IFUT *na*	*-xo-* *(ta)**	PRG *-a-*, *ra-***	CAUS *-vaxa-* REC *-e-* TOT *-va-*	[redup]		[verb semi-V noun

	RFUT *ba*	adverbs]
	CFUT *naba*	
	PRES/HAB Ø	

*_ta_ after _gi-_. Phonological interaction with tense _ta, ga, ba_ → _toxo_ etc.
**_a_ is present, _ra_ vague past.
Madak allows for only three markers to be present at any one time.

The Kuot verb phrase

| | | |

Appendix II. Pronominal forms

Forms which have already occurred in the same person-number function have been grayed.

Nalik pronominal forms

		indep.	S marker	DL: -a PC: -(mt)al	alien. poss.	inal. poss.
SG	1	*ni*	*ga*		*surago*	*-nagu, -nugu*
	2	*nu*	*gu*	+	*sunum*	*-num*
	3	*naan*	*ka, a, na*		*si-*	*-na*
NON	1n	*di*	*di(a)*	+	*si-*	*-di*
-SG	1x	*maam*	*madi*	+	*si-*	*-maam*
	2	*nim*	DL/PC**: *gu* PL: *nagu*	+	***	DL *-numa* PC *-numtal* PL *-nim*
	3	*na(a)nde*, *na(a)ndi*, *na(a)nda**	*di(a)*, (*ka*)	+	*si-*	*-naande*, *-naandi*, *-naanda*

*dialectal and generational variation.

**DL and PC are unusual; if used they have to go both on pronoun and V; 2nd person can have SG or NSG as base.

***DL in all persons by adding *si-* and *-(y)a* to PL inal; PC in all persons by adding *si-* and *-yal* or *-tal*.

Notsi pronominal forms

		S/indep.	O/prep_	alien. poss.	inal. poss.	ben-n-food	ben-food
SG	1	*ya*	*ya*	*nugu*	N-*g*V	*na-go*	*ma-ga-ka*
	2	*u*	*yu*	*nuum*	N-V*m*	*nuum-ka*	*maam-ka*
	3	*a-/Ø*	*Ø*	*kan*	N-*n*	*ka-ka*	*ma-ka*
DL	1n	*gita*	*gita*	*kida*	+ N-*n*	*mida*	–
	1x	*gelu*	*gelu-lu*	*na-gelu*	+ N-*n*	*na-gelu-ka*	*ma-gelu-ka*
	2	*gulu*	*gulu-lu*	*na-gulu*	+ N-*n*	*na-gulu-ka*	*ma-gulu-ka*
	3	*delu*	*delu-lu*	*ka-delu*	+ N-*n*	*ka-delu-ka*	*ma-delu-ka*
TL	1*n*	*gita-tuul*	*gita-tuul*	*kida-tuul*	+ N-*n*	*mida-tuul*	–
	1*x*	*getu*	*getu-tuul*	*na-getu*	+ N-*n*	*na-getu-ka*	*ma-getu-ka*
	2	*gutu*	*gutu-tuul*	*na-gutu*	+ N-*n*	*na-gutu-ka*	*ma-gutu-ka*
	3	*detu*	*detu-tuul*	*ka-detu*	+ N-*n*	*ka-detu-ka*	*ma-detu-ka*
PL	1n	*gita′*	*gita′*	*kida′*	+ N-*n*	*mida′*	–
	1x	*geem*	*geem*	*na-geem*	+ N-*n*	*na-geem-ka*	*ma-geem-ka*
	2	*gim*	*gim*	*na-gim*	+ N-*n*	*na-gim-ka*	*ma-gim-ka*
	3	*di*	*di*	*ka-di*	+ N-*n*	*ka-di-ka*	*ma-di-ka*

Inalienable possessives in non-singular numbers are formed in two ways:
1. alienable + possessed N + *-n*, or 2. possessed N + *-n* + object pronoun.

Madak pronominal forms

		Indep. S (*ne-*)	S marker	Indep. O	at- poss, prep:loc,tim	mi- comit, instr	ti- benef, goal etc	inal. poss.
SG	1	*nia*	*a*	*ia*	*-arak*	*-nia*	*-a*	*-k*
	2	*nu*	*u*	*u*	*-aram*	*-nu*	*-u*	*-m*
	3	*ni*	*i**	*i*	*-eren*	*-n*	*-n*	*-n (-no/-na/-on/)*
DL	1n	*da*	*ta*	*da*	*-da*	*-da*	*-da*	*-da*
	1x	*ma'*	*ma*	*nama*	*-nama*	*-nama*	*-nama*	*-nama*
	2	*mu*	*mu*	*numu*	*-numu*	*-numu*	*-numu*	*-numu*
	3	*du*	*du*	*du*	*-du*	*-du*	*-du*	*-du*
PL	1n	*dik*	*ta*	*dik*	*-dik*	*-dik*	*-dik*	*-dik*
	1x	*ma*	*ma*	*nama*	*-nama*	*-nama*	*-nama*	*-nama*
	2	*mi*	*mi*	*nimi*	*-nimi*	*-nimi*	*-nimi*	*-nimi*
	3	*di/ti*	*di***	*di*	*-di*	*-di*	*-di*	*-di*

*omitted before tense marker.
**omitted after *nedi*.

Kuot pronominal forms

num	pers	Pers pron	Dem; attr. relator	Subj aff Vcl.II,III, n-fut/fut	Inal poss, prep.	Obj pref; prep.	Adj.	Obj suff. Vcl.IIa 3rd, n-fut/fut	Subj encl., Vcl.I, n-fut/fut
SG	1	*turuo*	–	*tu/ta*	*tuo*	*to*	*to-* *-i*	–	*tuŋ/taŋ*
	2	*nunuo*	–	*nu/na*	*nuo*	*no*	*no-* *-i*	–	*nuŋ/naŋ*
	3m	–	*i*(-)	*u/a*	*a*	*a*	*-i*	*-a/-ŋ*	*oŋ/aŋ*
	3f	–	*u*(-)	*i*	*o*	*o* (*u*)	*-u*	*-o/-ŋ*	*ieŋ*

DL	1n	*bibi*	–	*bi*	*bi*	*bi*	*bi-*	*-n*	–	*biŋ*
	1x	*i*	–	*i*	*i*	*i*	*i-*	*-n*	–	*iŋ*
	2	*mame*	–	*ma*	*me*	*me*	*ma-*	*-n*	–	*maŋ*
	3	–	*li-*	*li*	*li*(*e*)	*le*		*-n*	*-an/-ŋan*	*lioŋ*
PL	1n	*bubuo*	–	*bu*	*buo*	*bu*	*bu-*	*-m*	–	*buoŋ*
	1x	*papa*	–	*pa*	*pa*	*pa*	*pa-*	*-m*	–	*paŋ*
	2	*mimi*	–	*mi*	*mi*	*mi*	*mi-*	*-m*	–	*miŋ*
	3	–	*mi*(-)	*me*	*ma*	*ma*		*-m*	*-am/-m*	*meŋ*

		Alienable possessives			
num	**pers**	**masc**	**fem**	**dual**	**plural**
SG	1	*tuaŋ*	*tuŋ*	*tuaŋan*	*tuam*
	2	*nuaŋ*	*nuŋ*	*nuaŋan*	*nuam*
	3m	*aŋ*		*aŋan*	*am*
	3f	*iaŋ*	*ieŋ*	*iaŋan*	*iam*
DL	1n	*biŋ*		*biŋan*	*bim*
	1x	*iŋ*		*iŋan*	*im*
	2	*meŋ*		*meŋan*	*mem*
	3	*liaŋ*	*lioŋ*	*liaŋan*	*liam*
PL	1n	*buaŋ*	*buoŋ*	*buaŋan*	*buam*
	1x	*paŋ*		*paŋan*	*pam*
	2	*miŋ*		*miŋan*	*mim*
	3	*meiaŋ*	*meioŋ*	*meiaŋan*	*meiam*

Appendix III. Kuot verbs and adjectives, and their affixation orders

Verb class		Intransitive	Transitive
I.	subj. enclitic obj. prefix	*pasei-oŋ* talk-3.M.SUBJ 'he talks'	*a-pasei-oŋ* 3.M.OBJ-talk-3.M.SUBJ 'he talks of him/it(m)' / 'he tells him'
IIa.	subj. prefix obj. suffix in 3	*u-libə* 3.M.SUBJ-cry 'he cries'	*u-alibə-o* 3.M.SUBJ-cry.over-3.F.OBJ 'he cries over it(f)/her [mourns]'
	obj. prefix in 1/2		*to-u-alibə* 1SG.OBJ-3.M.SUBJ-cry.over 'he cries over me'
IIb.	subj. prefix obj. prefix	*u-lo* 3.M.SUBJ-talk 'he talks'	*a-u-lo* 3.M.OBJ-3.M.SUBJ-talk 'he tells him'
III.	subj. "infix" obj. prefix	*uan-u-lə* wait-3.M.SUBJ-STM_2 'he waits'	*a-uan-u-lə* 3.M.OBJ-wait.for-3.M.SUBJ-STM_2 'he waits for him(/it.m)'
adj.	subj. suffix in 3	*mukə-u* pregnant-3.F 'she is pregnant'	–
	subj. prefix in 1/2	*to-mukə-i* 1SG-pregnant-SG 'I am pregnant'	–

Complexity in nominal plural allomorphy

A contrastive survey of ten Germanic languages

Antje Dammel
University of Mainz

Sebastian Kürschner
University of Groningen

We investigate the complexity of nominal plural allomorphy in ten Germanic languages from a contrastive and diachronic perspective. Focusing on *one* language family allows us to develop multidimensional criteria to measure morphological complexity and to compare different diachronical drifts. We introduce a three-step complexity metric, involving (1) a quantitative step, (2) a qualitative step, and (3) a validation step comparing the results from step (1) and (2) to actual language use. In this article, we apply the method's two first steps to the plural allomorphy of our sample languages. Our criteria include for (1) the number of allomorphs and for (2) iconicity in form-meaning relationship, the basis of allomorph assignment, and the direction of determination between stem and suffix. Our approach reveals Faroese as the most complex language and English as the simplest one.

1. Introduction

Number marking on nouns in Germanic languages reveals different degrees of complexity in formal marking.[1] English is usually regarded as belonging to the simple pole of a complexity scale because it has a small number of allomorphs which are mostly realized in suffix form. By contrast, Icelandic or Faroese are considered languages with a very complex system for number encoding because they have a large variety of allomorphs. Techniques used include stem alternation and zero marking in addition to different suffixes (cf. Wurzel 1990: 139 for an example of a morphological complexity scale with similar results).

Since the term 'allomorphy' is subject to a lot of different definitions, we specify this term for the sake of clarity. By allomorphs, we mean different formal realizations of

1. We would like to thank two anonymous reviewers and the editors of this volume for critical comments, which helped us to clarify important points in our discussion. We are grateful to Janet Duke for proofreading the manuscript in English. Of course, we are responsible for all remaining mistakes.

the same functional entity, i.e. the same morpheme. For the morpheme {noun plural}, German e.g., uses different suffixes, as *-n* in *Ente-n* 'ducks', or *-s* in *Auto-s* 'cars'. Not only suffixation is used to code the plural information, but also other coding techniques can be found. In *Kästen* 'boxes', e.g., the plural is coded by means of an alternation of the stem vowel, cf. the singular *Kasten*. We use the term 'coding technique' for the abstract formal procedure used, and the term 'allomorph' for the actual form used to signal the function in question. In this terminology, German *-s* and *-n* are examples of the technique of suffixation, but they are still two different allomorphs of the plural morpheme, as also the umlaut in *Kästen* is.

To achieve complexity measurements such as those for English and Icelandic, formal properties in morphological behaviour are usually taken as a basis for complexity judgements. Although this yields intuitively acceptable results, it often remains unclear what is meant by complexity, and for whom the systems are more or less complex. We define complexity as a system property which has effects on language users. We investigate complexity by contrasting different systems with respect to the plural formation of nouns. In Section 2, we present an analysis which combines the so called absolute and relative approaches to complexity.

We base complexity on several criteria in our account, including (a) the quantitative criterion of the number of allomorphs, and the qualitative criteria of (b) form-meaning relationships, (c) the assignment principles accounting for the appearance of allomorphs, and (d) the direction of determination between stem and suffix. We claim that one should consider these multidimensional criteria when one desires defining a sufficient basis for morphological complexity.

Our findings are based on a study of number allomorphy on nouns in the ten Germanic languages English (ENG), Dutch (DUT), Afrikaans (AFR), Danish (DAN), Swedish (SWE), West Frisian (FRI), German (GER), Icelandic (ICL), Luxembourgish (LUX), and Faroese (FAR) (Dammel, Kürschner and Nübling, manuscript). In Section 3, we examine and illustrate the multidimensional criteria considered for the ten languages. The results are then summarized in a complexity scale in Section 4. For our aim, the Germanic languages proved to be good test candidates since different typological drifts can be diachronically linked to their common ancestor language.

Our scope is purely morphological and we have only taken nouns into account. We left out the nominal phrase (NP) as a larger domain for plural encoding, even though it could have provided clues for a more exact metric of complexity. In future research, larger domains such as this should be integrated into the metric proposed here.

2. Defining and analysing morphological complexity

With regard to the notion of linguistic or – more specifically – morphological complexity, different complexity metrics have been developed in recent years. In order to distinguish our notion of complexity from other notions, we need to specify what

we mean by this term. The main idea in our definition of complexity is that language systems are characterized by a certain degree of complexity which can be evaluated in comparison with other language systems. A point of importance is that we define complexity as a property of the language system which is analyzable with respect to its effect on language users. We agree with e.g., Hawkins (2004) in that language systems are not totally independent of performance principles. System complexity can only be validated by accounting for the difficulty a system causes to its users.

We define a way to measure *morphological* complexity. Thus, we only investigate *one* level of the language system. We wish to state clearly that we do not mean to measure overall *linguistic complexity*. We chose the domain of nominal plural allomorphy as a testing ground for our metric, which can be applied to other morphological domains in future work. Because all of the ten Germanic languages contrasted here possess allomorphic representations for plural information, we can be sure to investigate *system complexity* and not *expressivity* in Dahl's (2004: 43) terminology. Dahl differentiates between the resources of a language (e.g., the number of grammatical functions coded, which in his definition does not belong to the measurement of linguistic complexity, but of the "richness" or expressivity of a system) from "the system of regulations that determines how to express that which can be expressed."

Miestamo (2006, this volume) contrasts "absolute" and "relative" approaches to investigating linguistic complexity. Absolute approaches such as McWhorter (2001) and Dahl (2004) view complexity as an objective system property whereas relative approaches base complexity on the difficulty a language causes to its users (Kusters (2003: 6) defines complexity as "the amount of effort an outsider has to make to become acquainted with the language in question," an outsider being a second language learner).

Our account provides a combination of these different approaches, using the advantages of both. For this reason we propose a three-step metric: In the first step, we count the number of regular forms for plural expression, i.e., the number of regular allomorphs – and thereby plural classes. In the comparison of different systems, the higher the number of allomorphs is, the higher the complexity of the actual system. With this method, e.g., Swedish – with five productive plural allomorphs (*-ar, -er, -or, -n,* and zero) – has a higher degree of morphological complexity than Dutch, which has two productive allomorphs (*-en,* and *-s*). This quantitative step resembles absolute approaches, since it bases complexity on the differing number of formal instances.

However, a quantitative account such as this is not applicable to all aspects of plural formation. Since all our sample languages use more than one technique to express plural information (e.g., zero marking or stem alternation in addition to suffixation), a measure for the complexity of these different techniques is necessary. Again, with regard to the first step, we could do this by counting the different techniques.

Nevertheless, the complexity of coding techniques can also be determined by means of the effects they cause on lexicon organization or the (de-)coding of information. These questions cannot be completely solved in absolute terms, but only in relation to the effect they have in speech-/perception processes. In order to come to grips with

qualitative complexity criteria such as these, we examined morphological theories which model linguistic procedures in the mind. We base our second step on considerations from morphological theories (cf. Section 3.2.1) and develop theoretical statements from the data in question. Although the morphological theories we use aim in themselves to be objective models, we doubt that this step represents an absolute approach to complexity. In contrast, the considered theories use models of cognitive language processing. This means that theories built on models of cognition are based on relative notions (see the discussion in Kusters, this volume). On the other hand, theoretical assumptions are not relative in the sense that language use is the starting point of the investigation of complexity (as in Kusters' own relative approach).

According to these considerations we believe that a third notion between absolute and relative approaches is necessary. In this step, a morphological theory offers a cognitive model of language use to base notions of complexity on. We can call this step a "model-based relative" approach to complexity, because the modelled difficulties in language use, which are taken as a basis for complexity considerations, still retain a hypothetical character. Validation in actual language use remains to be done.

Although in this article we can only deal with the two steps introduced until now, we propose a third step in our complexity metric which is not only a "model-based relative" step but an "actual relative" step. It should consider difficulties experienced by different language user groups (e.g., data from experiments on native speakers' language processing) or by language learner groups (first and second language acquisition by children and adults), and draw on results of psycholinguistic experiments (Section 5). The results of the first two steps should be validated with these data.

To sum up, the metric of morphological complexity proposed in this article is based on two steps: (1) *a quantitative account* of the forms available and (2) *a qualitative account* based on morphological theories. For future work, we propose a third step, (3) *a validation process* comparing the results with data about the complexity of actual language use.

Before starting the actual application of our metric, we would like to state that our notion of morphological complexity should not be misunderstood to be a judgement about a language's value. We understand complexity simply as a system property, which is totally unrelated to any social judgements.

3. Illustration of the complexity metric

This section illustrates how our complexity metric works for the first two steps described in Section 2. In the first step, the ten languages are ranked according to the quantitative criterion number of plural allomorphs (3.1). In the second step, we consult qualitative criteria (3.2), considering the following three dimensions: iconicity of the formal techniques used (3.2.1), assignment principles (3.2.2), and direction of determination between stem and plural marker (3.2.3). We provide complexity

rankings for each dimension. We examined data from secondary sources such as standard grammars, dictionaries, and surveys of plural formation in single languages.[2]

3.1 Number of plural allomorphs across Germanic languages

In this section we describe how we counted the number of allomorphs and used the number for the complexity metric. For the following inventory (see Figure 1), we considered only indigenous, type frequent and/or productive plural allomorphs.[3] Different vowel alternations were not rated separately, but rather subsumed under the overall technique. Had we proceeded differently, the number of allomorphs of those languages with high amounts of stem alternation – German, Icelandic, Faroese, and especially Luxembourgish – would have increased remarkably. Nübling (2006: 110) counts about 50 allomorphs for Luxembourgish and 28 for German in that way. This shows that even a supposedly strict criterion for complexity, such as number of allomorphs, depends to a certain degree on the analyst's judgments.

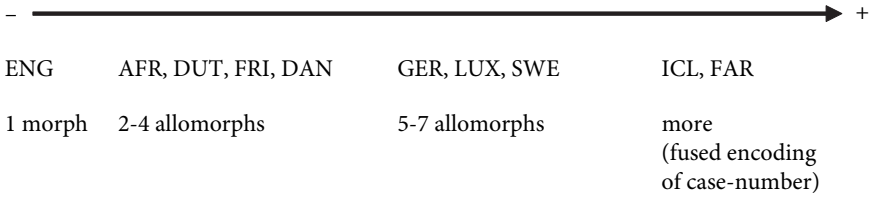


Figure 1. Amount of allomorphy across Germanic languages.

Figure 1 shows English as the simplest language with only one suffixal morph (with the variants -[s], -[z], and -[ɪz]), and then two medium groups: Afrikaans, Dutch, Frisian and Danish with two to four suffixal allomorphs, and German, Luxembourgish and Swedish with five to seven. While Swedish relies mainly on suffixes, German and Luxembourgish use and combine suffixation as well as stem modulation. Icelandic and

2. Exemplary sources are for Afrikaans: Donaldson (1994), Raidt (1983); Danish: Allan, Holmes and Lundskær-Nielsen (1998); Dutch: Haeseryn et al. (1997); English: Quirk et al. (1985); Faroese: Thráinsson et al. (2004); German: Augst (1975), Duden (2005), Eisenberg (2004), Neef (2000b); Icelandic: Kress (1982); Luxembourgish: Christophory (2004), Derrmann-Loutsch (2004), Keller (1961); Swedish: Holmes and Hinchliffe (1994); West Frisian: Markey (1981), Tiersma (1999).

3. With respect to irregularity, it seems more appropriate to count the number of nouns with irregular plural forms: regular mechanisms can be defined in plural classes, so every noun needs to be learned with the class (defined as regular morphological behaviour) it belongs to. Irregular plural formation, on the other hand, is idiosyncratic in the sense that a word needs to be learned without a class, but together with its morphological forms (see stem involvement, especially suppletion, in 3.2.1 and idiosyncratic allomorph assignment in 3.2.2). These are two different criteria of complexity.

Faroese show the highest amount of allomorphy because they have retained fused encoding of number and case. Thus, we rank them as most complex.

3.2 Three dimensions of complexity in plural exponency and assignment

3.2.1 *Complexity of formal techniques*

Complexity decisions for this section are based on traditional accounts such as Markedness Theory and Natural Morphology (see e.g., Dressler et al. 1987; Wurzel 1998) as well as Bybee's (1985) principle of relevance. The main point is that they rely on an ideal of iconicity in form-meaning relationship. In Natural Morphology, the form-meaning relation should be one-to-one. An addition in meaning should yield an addition in form (termed constructional iconism by Natural Morphologists). This favours material expression of number, and disfavours syncretism and zero expression but also allomorphy and redundant marking. Secondly, formal exponents should be clearly segmentable. This favours agglutinative, concatenative exponence and disfavours fused and/or modulative (e.g., umlaut) exponence.

The iconicity in Bybee's principle of relevance is more intricate. In our context, it refers to the impact the semantics of the grammatical category number has on the semantics of the word class noun it attaches to. Thus, number information, which replicates the concept, modifies the semantics of the noun more than e.g., case, which determines only its syntactic role (cf. Bybee 1994: 2559).[4] According to Bybee, the degree of relevance is mirrored formally in the degree of fusion. Thus, we claim in our context that the relevant number information should attach to the noun itself and favour inflectional rather than syntactic expression more than e.g., case marking does.

These theoretical claims yield material number marking by suffixation and without allomorphy as the canonical simple pole of our formal complexity scale, because it is easily segmentable and meets the ideal of constructional iconism: the addition in meaning, in this case plural information, is mirrored in form by an addition of material to the stem.

In this sense, we define formal complexity in terms of iconicity in a form-meaning relationship. Formal complexity comprises all deviations from an iconic one-to-one-relationship (for a similar approach see Kusters' principle of transparency (2003: 24–26)). We consider the following aspects of this relationship: (1) stem involvement, (2) redundant marking: combination of techniques, (3) zero expression, (4) subtractive expression, (5) allomorphy, and (6) fusion of number and case expression.

3.2.1.1 *Stem involvement*. We consider a high degree of stem involvement as complex; it represents additional encoding effort and transparency diminishes because the

4. Booij's (1996) concept of inherent vs. contextual categories could also have been used here.

plural information becomes hardly segmentable. A high degree of stem involvement correlates with the plural forms' lexicalization and thus irregularity.[5]

The degree of stem involvement increases from right to left. Thus, it is higher if an alternation affects the stem vowel or even the consonantal onset than if it is constrained to the stem final consonant. Suffixation, which is the main device in all sample languages and almost the only technique in English, is the least complex as it implies no stem change at all. Next in complexity is stem alternation, which we subdivide into several complexity degrees. Alternation of stem final C, such as consonant sonorization in Dutch, Afrikaans, Luxembourgish (and as a relic in English, e.g., *house* -[s] – *houses* -[zɪz], *dwarf* -[f] – *dwarves* -[vz]) are considered as less complex than alternations of the stem vowel, because the former are generally more predictable and often contextually triggered. In Luxembourgish for example, plosives as well as fricatives are devoiced word finally, but voiced in onset position. This leads to consonant alternations between singular and plural such as e.g., [s] – [z] in *Glas – Glieser* 'glas(ses)', and [f] – [v] in *Bréif – Bréiwer* 'letter'. Faroese alternation of stem final consonants is more complex because phonological and morphological processes interact. Among stem vowel alternations, qualitative alternations such as umlaut in German, Luxembourgish, Icelandic, and Faroese are classified as more complex than quantitative ones (e.g., vowel lengthening in Danish *blad* [a] – *blade* [aː]), because more features of the sound are modified. Arbitrary patterns are considered more complex than transparent ones. The former is the case in Luxembourgish; compare e.g., Luxembourgish *Toun – Téin* [ou] – [ɛɪ] 'tone', *Wuert – Wierder* [uə] – [iə] 'word' with transparent German palatalization in *Ton – Töne* and *Wort – Wörter*. Stem alternation has morphologized, i.e., has become productive for number marking only in German and Luxembourgish. Here, it spread e.g., to masculine *a*-stems such as German *Stab – Stäbe* 'staff' and neuter *a*-stems such as German *Wort*/Luxembourgish *Wuert*. In Luxembourgish umlaut functionalization was most intense; even originally non-umlautable stems are affected, e.g., *Rass – Rëss* [a] – [ə] 'crack(s)', the stem vowel of which goes back to Middle High German /i/ (for further details see Nübling 2006: 116f.). Suppletion, which is also a gradient notion, is the most complex type of stem involvement. Partial suppletion is more frequent in Luxembourgish and Faroese than in other languages, e.g., Luxembourgish *Steen – Steng* ['ʃteːn] – ['ʃtɛŋ] 'stone' or Faroese *dagur* ['dɛavʊr] – *dagar* ['dɛːar] 'day(s)'. Total suppletion is mainly restricted to words for professions such as *businessperson – businesspeople* (German *-mann/frau – -leute*, in several other languages as well).

5. Nevertheless, stem involvement does not automatically imply irregularity, as e.g., regular German or Icelandic umlaut, or completely regular sonority neutralization in final consonants of Dutch and Luxembourgish show. Therefore we treat the notions stem involvement and irregularity separately.

In order to evaluate stem involvement, we ranked those techniques found in the Germanic languages left to right on a scale in Table 1 according to their increasing degree to which the stem is affected and beginning with the stem's right periphery. We ranked the languages vertically from top to bottom by their increasing degree of stem involvement. The number of plus signs outlines the amount of use of the respective technique in a language system. We fixed it according to the data regarding frequency and productivity stated in our sources (see note 2). Thus, the plus signs are no exact quantification (which has yet to be developed) but merely a rough illustration. We are aware of the problems this implies, but still, this illustration offers an easily accessible overview of the empirical facts.

Table 1. Degrees of stem involvement across Germanic languages.

Language	Suffixation without stem modulation	Stem modulation			Suppletion (degrees)	Form-meaning relation
		final C	root V			
			quant.	qual.		
	←				→	– redundancy and allomorphy +
ENG	++++	(+)		(+)		
DUT	++++	+	+	(+)		
AFR	++++	+	+	(+)		
DAN	++++		(+)	+		
SWE	++++			+(+)		
FRI	++++		+	++		
GER	++++	+		++(+)		
ICL	++++			+++		
LUX	++++	+	++	+++	+	
FAR	++++	+++	++	+++	++	

Stem involvement often correlates with or implies redundancy (3.2.1.2), e.g., German *Ton – Töne*, or the partially suppletive Luxembourgish example *Steen – Steng* above. However, both criteria should be kept distinct, as applying one single non-concatenative technique alone is not redundant, compare e.g., non-redundant Luxembourgish *Toun – Téin*. The same holds for the correlation between stem involvement and allomorphy (3.2.1.5) (see rightmost box in Table 1).

3.2.1.2 *Redundant marking: Combination of techniques*. We regard redundant marking only as far as the noun itself is concerned and generally blend out other NP constituents. Redundancy violates iconicity, as several formal techniques express only one meaning. A combination of different (concatenative and modulative) techniques occurs frequently in German and Luxembourgish (see above for examples), Faroese and Icelandic, leading to noun-internal redundant marking (e.g., Icelandic *fjörður – firðir* masculine 'fiord', *höfn – hafnir* feminine 'port').

The problem in measuring the complexity of redundant marking is that it implies a coding-decoding asymmetry: more speaker effort, but additional decoding cues for the listener. This is a relative argument. In an absolute approach, redundancy would be regarded complex, as it requires a lengthier description. But is length of description

really symmetrical for input (length of description) and output direction (how early is ambiguity solved) – aspects which even an absolute approach should consider?

Diachronically, redundancy by combination of techniques can be a step towards regularization, e.g., when a productive suffix is added to an unproductive one. This happened e.g., in Dutch *kind-er-en*, where productive *-en* was appended to unproductive *-er*.

We regard redundancy as complex simply on account of the theory we used (Natural Morphology). The validation of this theory-based decision remains to be done in future research in the third step of our approach (see Section 5).

3.2.1.3 *Zero marking*. Zero marking occurs productively and with a high type frequency only in North Germanic languages (mostly for neuter gender), e.g., Danish *et <u>år</u>* – *mange <u>år</u>* 'a year – many years', in German (for masculine and neuter nouns ending in a reductive syllable), e.g., *der <u>Löffel</u>* – *die <u>Löffel</u>* 'the spoon(s)', and to a lesser extent in Luxembourgish (see Table 2).

Determining the complexity of zero marking again reveals the relativity of qualitative criteria towards complexity. From a decoding perspective, zero is quite complex, but from encoding perspective it seems at first very simple. However, on theoretical grounds we rank it as complex: zero marking violates one-function-one-form and can lead to ambiguity as nouns are the only obligatory constituents of the NP. Even though dependent marking often solves this ambiguity, on theoretical grounds (cf. Bybee's relevance principle introduced above) the inherent grammatical category number, which is highly relevant to the noun, should rather be marked on the head itself more than e.g., case should. For now, the additional consideration of dependent marking remains an unresolved issue, but a preliminary statement could be that zero marking, which is not compensated for at all, is more complex than that compensated by dependent marking.[6] For example, in phrases such as the following, zero marking is not compensated for: Icelandic *Hanna sér borð* 'Hanna sees a table/tables' or Luxembourgish *Si gesäit d'Millen* 'She sees the mill(s)'.

3.2.1.4 *Subtractive marking*. Subtractive marking occurs exceptionally in Luxembourgish (and German Franconian dialects) due to intervocalic consonant

6. However, cross linguistic evidence provides no cue whether head or dependent marking is more complex (cf. Nichols 1992: 69–76 for a discussion). The editors and one of the reviewers posed interesting questions regarding the complexity of head and dependent marking which we cannot solve here but nevertheless want to mention: is, on the one hand, double marking on head and determiner - being redundant - more complex than simply on one of them? Or is, on the other hand, a lack of agreement between both (either head marking or dependent marking only) to be regarded as complex? This would also be interesting to discuss from an absolute point of view.

assimilation followed by apocope of the triggering vowel, which leads to e.g., Luxembourgish *Frënd – Frënn* (< **Frënne* < **Frënde*). We regard this device as complex because it is countericonic (see Table 2).

3.2.1.5 *Allomorphy*. Allomorphy, which means several choices of form for one and the same function and thereby a deviation from uniform encoding, is considered complex. The amount of allomorphy was already sketched in Section 3.1 (Figure 1) and is outlined again for convenience in Table 2 with Icelandic and Faroese as the most complex and English as the simplest system. Allomorphy occurs twice in our survey, as it is open to examination from both a quantitative (3.1) and a qualitative point of view (Section 3.2.1).

3.2.1.6 *Fused encoding of number and case* Fused encoding is complex as it diminishes transparency and segmentability and increases allomorphy. It was historically retained only in Icelandic and Faroese (see Table 2), both for suffixation and stem alternation. The plural suffixes of Icelandic *hestur* 'horse' e.g., are always fused with case information:

(1) *hest-ar* *hest-a* *hest-um*
horse-NOM.PL horse-GEN/ACC.PL horse-DAT.PL

In Faroese *brúgv* 'bridge' this even correlates with stem alternation:

(2) *brýr* *brúa* *brúm*
[brɯʏr] [brʏua] [brʏun]
bridge.NOM/ACC.PL bridge.GEN.PL bridge.DAT.PL

Table 2. Formal complexity: deviations from iconicity in form-meaning relationship.

Complexity	Stem involved	Redundancy	Zero	Subtraction	Allomorphy	Fusion number/case
–	ENG			Others	ENG	Others
	DUT					
	AFR	Others	Others			
	DAN					
	SWE				AFR,	
	FRI		LUX		DUT, FRI,	
					DAN	
	GER					
	ICL	LUX,	DAN,	LUX	SWE,	
	LUX	GER, ICL,	SWE,		LUX,	
		FAR	FAR, ICL,		GER	
			GER			
+	FAR				ICL, FAR	ICL, FAR

3.2.2 *Complexity in allomorph assignment*

Allomorphs are regularly assigned to stems according to specific structural properties of the stem, e.g., the final sound or the stem's gender (cf. Neef 2000a, b). The principles at work in allomorph assignment may be based on different grammatical layers – either

aspects of phonological form (signifiant) or aspects of grammatical and lexical function or content (signifié) (see 3.2.2.1). They may act in isolation or in interplay with other assignment principles (see 3.2.2.2).

3.2.2.1 *Formal, functional, and idiosyncratic assignment*. We arranged the most common formal and functional assignment principles found in the Germanic languages on a complexity scale based on the "lexical depth" of the layers involved (see Table 3). This means that assignment relying on formal factors is less complex than assignment drawing on the content part of the lexical entry. Idiosyncratic assignment is not subject to any regularity at all, but has to be specified in the lexical entry. It is therefore the most complex.

Thus, phonologically conditioned assignment is the least complex, as it draws exclusively on formal features such as final sound or prosodic pattern, and does not access the content part of the lexical entry. Internally, prosodic factors are more complex than segmental ones such as final sound, because they draw on the stem as a whole (cf. Dahl 2004: 51–55 for the non-linearity and thus complexity of prosodic properties). Next in complexity is morphologically-conditioned assignment, e.g., by derivational suffixes, which draws on form as well as function. In the Germanic languages this factor is generally dominated by gender and/or phonological criteria and will therefore be neglected in the following. Assignment based on semantic factors (e.g., animacy) is more complex than morphological assignment, as no formal cues remain. Assignment conditioned by gender is even more complex, for there is scarcely any semantic bridge left as a mnemonic aid in the highly grammaticalized gender distinctions of the Germanic languages. Nevertheless, gender is more regular than idiosyncratic assignment, because all nouns share the property of belonging to a gender and thus form groups, which manifest themselves in agreement. In this way, the lack of any regular principle at all, i.e., idiosyncratic assignment, implies the highest degree of complexity, as inflectional forms here are stored uneconomically as separate, holistic units which are not subject to general rules. Consider e.g., the different plural exponents German *Hunde*, *Münder* and *Gründe* for the masculine rhyming triple *Hund*, *Mund*, *Grund* 'dog, mouth, reason'.

English is a well-known example of a system depending wholly on final sound, which is the most formalized of all assignment principles. The only suffix {-S} has three variants, which assimilate to the final sound of the stem: -[ɪz] after sibilants, e.g., *faces*, -[s] after voiceless non-sibilants, e.g., *bats*, and -[z] after voiced non-sibilants, e.g., *frogs*. Frisian, Dutch and even Afrikaans are more complex, as they combine two criteria, final sound and prosody. In Dutch e.g., nouns with final stress and disyllabic nouns ending in *-e* such as *stoel – stoelen* 'chair' and *blinde – blinden* 'blind person' take *-(e)n* plural. Disyllabic nouns ending in *-e* plus sonorant such as *vogel – vogels* 'bird' take *-s* plural (the latter is currently spreading to further contexts).

Semantic assignment occurs in most languages irregularly for small groups such as inherently pluralic concepts. In several languages, zero plural was kept or even extended for animals/plants occurring in groups or pairs such as e.g., English *sheep*, *deer*, *fish, lion, antelope* or Luxembourgish *Schof* 'sheep', *Bier* 'berry', *Been* 'leg', *Fësch* 'fish'.

Danish, and to a lesser extent German, even use semantic assignment regularly. Danish, though having generally formalized its system, developed animacy as an assignment principle for the *e*-plural which may overrule derivative criteria: the suffix *-ing* usually triggers *-er*-plural as in *blanding – blandinger* 'mixture(s)', but with animates it takes *-e*. Thus, e.g., *udlænding* 'foreigner' is pluralized as *udlændinge*. In Danish, gender reduction leading to the merger of the masculine and feminine genders seems to have paved the way for strengthening the semantic factor of animacy. In German, animacy determines assignment to the weak masculine stems (thus being subordinated to gender): for weak masculine nouns (*n*-stems), e.g., *Bär – Bären* 'bear(s)', animacy (combined with formal factors) developed into a main assignment principle. While inanimates were weeded out, this plural class is still productive for highly animate concepts such as nationality terms, e.g., *Finne – Finnen* 'male person(s) from Finland' (cf. Köpcke 2000).

An illustrative example for gender-dominated assignment is found in the Swedish system: the suffixes *-ar*, *-(e)r* and *-or* are largely restricted to common gender nouns such as *bil – bilar* 'car(s)', *film – filmer* 'film(s)' and *flicka – flickor* 'girl(s)'. The suffix *-n* and zero marking are almost exclusively restricted to neuters, e.g., *bi – bin* 'bee(s)' and *ägg – ägg* 'egg(s)'. Gender is also dominant in the plural assignment of Icelandic, Faroese (to a lesser extent than in Icelandic), Luxembourgish, and German. Except for Swedish, this list comprises only those languages which have not undergone gender reduction (cf. Table 3 and Figure 2, bold print).

A high amount of idiosyncratic, and thus, completely lexicalized assignment increases complexity. It occurs especially in those languages with function-based assignment, e.g., Luxembourgish and Icelandic (see Table 3).

In Table 3 we ranked the languages according to the factors introduced in this section. The number of plus signs outlines the degree to which the respective criteria (scaled horizontally) prevail in the assignment systems of the languages (ranked vertically) according to our sources (cf. note 2). Again, this is merely meant as an illustration, not as an exact quantification.

Table 3. Complexity scale of assignment principles.

Language	Form-based		Function based		Idio-syncratic
	Final sound	Prosody	Semantics	Gender	
ENG	+++++				+
DUT	++	+++			+
FRI	++	+++			+
AFR	+	++++			+
DAN	+	++	++		+
SWE	++	+		++	+
FAR	++	++		++(+)	+++
GER	++	++	(+)	+++	++(+)
LUX	++	++		+++	+++
ICL	++	++		+++	+++

A comparison of Table 3 and Figure 2 shows that gender reduction correlates strongly with formalization (and in Danish with semantization) of plural assignment. Those languages which maintained a three-gender system retained or even strengthened gender as a main assignment principle. The breaking point towards more formal assignment seems to be a two-gender system (underlined in Figure 2). The genderless systems show the most formal assignment (italics). However, Swedish and Faroese violate this correlation: Swedish kept up gender in assignment despite gender reduction from three to two, Faroese weakened gender in assignment slightly, even though it retained three genders.

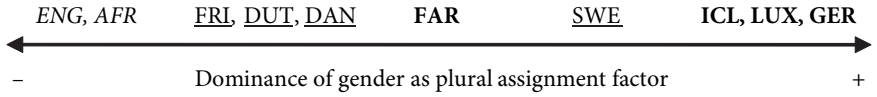


Figure 2. Correlation between gender as an assignment factor and gender maintenance (boldface: three, underlined: two, italics: no genders).

3.2.2.2 *Interaction of assignment principles.* Table 3 also provides an impression of the extent to which different principles occur in combination and interact in a language system. If assignment principles combine and/or interact to a high degree, this yields an increase in complexity for the respective system. Combining principles is again typical for those languages with function-based assignment. An example to the point is German, which relies generally on a combination of gender, accent pattern and final sound (cf. Neef 2000b: 479–480), and in the special case of weak masculines (see above) additionally on the semantic factor animacy, while e.g., English and Dutch employ phonological assignment throughout.

3.2.3 *Direction of determination between stem and suffix*

The third and final dimension relates to the direction of determination of formal features between stem and suffix (see Figure 3). In languages with right-to-left direction the suffix determines formal features of the stem, e.g., the German *-er* suffix triggers umlaut alternation (as in *Haus* [aʊ] – *Häuser* [ɔɪ] 'house(s)'). This direction is inherited from the earliest stages of the Germanic languages. It was retained by those languages with function-based assignment (see Tables 1 and 3). We have discussed its level of complexity under the label of stem involvement (3.2.1.1).

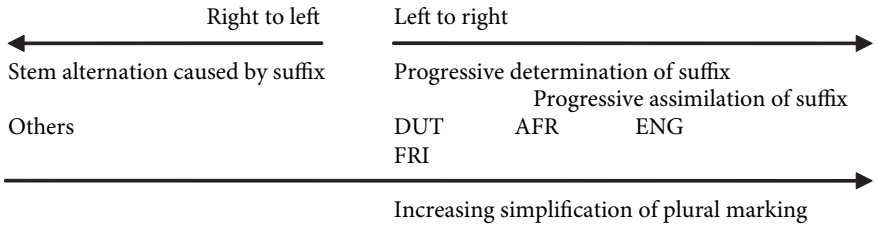


Figure 3. Direction of influence between stem and suffix.

Especially those languages with form-based assignment levelled out stem modulation (see Tables 1 and 3) and instead developed and/or strengthened its counterpart,

left-to-right direction. Here, formal features of the stem determine the choice or even the shape of the suffix. We regard this diametrical change of direction compared to Proto-Germanic as a further indicator of formalization and, as such, simplification. The weaker form of left-to-right direction is left-to-right determination, where phonological features of the stem – such as accent pattern and final sound – determine the choice of suffix. This has been discussed above (3.2.2) as formal assignment (e.g., in Dutch and Frisian). English goes a step further, as the suffix is even assimilated progressively by the stem. This is the summit of formalization and thus, simplification. Afrikaans holds a medium position: it shows mainly left-to-right determination, but it also developed a new plural suffix *-te* by way of left-to-right reanalysis of the stem-suffix boundary based on the stem's phonological structure (see Table 4).

Table 4. Development of Afrikaans *-te* plural through progressive reanalysis of the stem-suffix boundary.

	'writing, script etc.'	**'guest'**
Sg. Simplification in coda:	**skrift > skrif*	**gast > gas*
Pl. Preservation in onset:	*skrif.t-e*	*gas.t-e*
Reanalysis of morpheme border:	*{skrif.t}-{e} > {skrif}-{te};*	*{gas.t}-{e}> {gas}-{te}*
Secondary extension to nouns ending in *-s*:	*bos – boste* 'bush', *tas – taste* 'cup', *mens – menste* 'man'	

We regard this as a left-to-right influence which is stronger than determination but weaker than assimilation.

For the sake of clarity, we introduced and illustrated each of our multidimensional complexity criteria in isolation here, and provided rankings of the ten languages considered for each criterion separately. In Section 4, we will sum up these results in a survey of overall complexity of plural morphology.

4. Results

Various criteria concerning the morphological complexity of number marking were introduced and illustrated in Section 3. From a contrastive point of view, we can now rank the languages according to the results achieved for each complexity criteria. We arranged the languages into a summarizing chart (Table 5). For each complexity criteria (arranged in the columns) the languages are ordered vertically on a complexity scale, with a simple pole on the top, a complex pole at the bottom, and intermediate positions in between. When the number of allomorphs is considered, for example, we distinguish between four categories (illustrated in Section 3.1): English – with only one allomorph – is found at the simple pole, Icelandic and Faroese at the complex pole, and in between there is a simpler section of Afrikaans, Dutch, Frisian, and Danish on the one hand, and a more complex section of Swedish, German, and Luxembourgish on the other. All languages with equal complexity are grouped and separated by commas.

Table 5. Morphological complexity in contrast – a diffusion chart.

Complexity	Number of allomorphs	Formal techniques					Assignment	Direction of determination	Amount of regularity
		Stem involvement	Redundancy	Zero marking	Subtraction	Fusion of number/case			
simple pole	*ENG*	*ENG*	*FRI, DUT, AFR, ENG,* SWE, DAN	*ENG, FRI, DUT, AFR,* DAN, SWE, GER, **FAR, ICL**	*ENG, FRI, DUT, AFR,* DAN, SWE, GER, **LUX**		*ENG*	*ENG*	*ENG*
–		*AFR, DUT*		**LUX**			*DUT, FRI*		*AFR, FRI*
	AFR, DUT, FRI, DAN	DAN, SWE						*AFR*	*DUT*
				DAN			*AFR*		DAN
		FRI							
							DAN		SWE
	SWE, GER,	GER					SWE	*DUT, FRI*	GER
+	**LUX**	**ICL**		SWE			**FAR**		**LUX**
		LUX							**ICL**
complex pole	**ICL, FAR**	**FAR**	**LUX**, GER, **FAR, ICL**	GER, **FAR, ICL**	**LUX**	**ICL, FAR**	GER, **LUX, ICL**	**FAR, ICL, LUX**, GER, SWE, DAN	**FAR**

We only indicate the necessary degrees of complexity. Therefore, we distinguish only two complexity grades for redundancy, subtraction, and fusion, while there are more (finely-grained) degrees for the other criteria (e.g., stem involvement or assignment).

We represent formal techniques in five subcategories. Furthermore, in addition to the criteria illustrated in Section 3, we include the amount of regularity in the single languages' lexicon. This criterion is an implicit combination of the amount of suppletion and the amount of idiosyncratic assignment (see Section 3). The lower the amount of regular nouns in the lexicon is, the more complex we consider the morphology of the language in question to be.

In sum, we evaluate the languages according to nine complexity criteria. Through the analysis of their horizontal diffusion in the chart, it is now possible to determine the complexity status of each language. Afrikaans, for example, always occurs at the simple pole or very close to it, and is therefore characterized by low morphological complexity. The same holds for Dutch and Frisian, even though they tend to be a bit more complex in comparison with Afrikaans. English is the only language which is always placed at the simplest pole, and we can therefore conclude that its plural allomorphy is the least complex. We indicate the four languages close to the simple pole by italics.

Faroese is found seven out of nine times at the complex pole; its plural allomorphy is therefore the most complex of the languages examined. Icelandic and Luxembourgish are also often found at the complex pole or very close to it. We indicate the complexity of these three languages by bold typeset. The three remaining languages, Danish, Swedish, and German, appear mainly in between the poles. Even though German seems to be more complex than Swedish, which, in turn, seems more complex than Danish, these three languages have an intermediate status, oscillating between complexity and simplicity. In this way, three groups can be identified from the diffusion of complexity criteria: languages with complex (bold print), simple (italics), and intermediate plural allomorphy (normal print).

It is obvious, that most of the languages do not have a straightforward complexity status with regard to all the criteria. For example, Luxembourgish is found close to the simple pole for zero marking, and directly at the simple pole for fusion of number and case, although overall it tends to occur at the complex pole. The same holds for Faroese and Icelandic when we take subtraction into account. This shows how useful it is to examine a multidimensional range of criteria and to determine the languages' complexity by taking their diffusion regarding all these criteria into account. In this way, one may consider further criteria, for example by adding the surrounding NP or case allomorphy to the domain of examination. The results would probably become even more reliable by including further criteria: for example, when case allomorphy is considered, Faroese is less complex than Icelandic, because case marking on nouns was reduced from four to three cases. In our analysis a fact like this could not appear because we restricted our domain of investigation to number allomorphy and left out the case system due to methodological reasons.

By drawing our results from the diffusion with regard to a number of criteria, we address a serious problem of complexity research, which Miestamo (2006, this volume)

calls "the problem of comparability". According to this problem, there are no objective means to determine how much influence single criteria in a complexity metric have on the overall complexity, i.e., to determine the criteria's relative weight. By employing the diffusion with regard to our criteria, our analysis gets by the problem of comparability by giving the considered criteria equal weight. Overall results can be drawn from the languages' behaviour with respect to a larger number of criteria, so decisions about the weight of single criteria become unnecessary. Still, our results can only be interpreted as tendencies rather than absolute statements.

As indicated several times in Section 3, our results allow for some conclusions about pathways of morphological (number) simplification from a diachronic point of view. Morphological simplification proceeds not only by formal means, but also includes regular steps in the development of assignment principles: while languages that keep up the gender distinctions inherited from Proto-Germanic are complex regarding assignment, gender simplification brings about formalization in allomorph assignment, and sometimes even strengthens semantic factors, as in Danish. Furthermore, we can draw conclusions about the simplification in terms of the direction of determination: again, the complex languages keep the form of right-to-left determination common to all of the earliest Germanic languages, which results in high degrees of stem involvement for morphological matters. Simplification here corresponds with left-to-right determination (e.g., Dutch), and in English even the state of suffix assimilation was reached. At the same time, regularity in the lexicon seems to increase while allomorphy diminishes. It is interesting that languages with an intermediate status are found along these pathways according to some of the criteria, but still retain complexity at certain other points. This could be interpreted as a transitional stage for these languages.

5. Remarks on validation and conclusion

In our analysis we examined complexity according to the first two steps of a three-step metric. Still, we believe that complexity can only be a relevant linguistic notion when it affects the actual users of the examined systems. The results achieved in the first two steps therefore need to be validated by data from language use, which should bring about a relative component by including the persons to whom the systems actually are (or are not) complex. This third step, which we call "validation", remains to be investigated in further detail.

In this article, we can only present notions of our future work in very general means. When validating, it will be necessary to consider two general methodological problems: (1) the number of languages we examine is high, and (2) there is a large number of user groups which could be considered.

Psycholinguistic studies of inflectional morphology on nouns have not been conducted for all of the languages examined. Although psycholinguistics has developed rapidly in recent years for English and German as well as Dutch, the smaller languages such as Faroese and Luxembourgish have not been subject to psycholinguistic studies

to any great extent. Therefore, results from other studies will need to be generalized, which keeps the results vague. With respect to different user groups, psycholinguistic research is carried out e.g., on first and second language acquisition as well as on native speakers using their mother tongue. Results can differ according to the group investigated (cf. Kusters 2003: 36–41).

In the following, we provide some examples of validation. These are very general examples and the methodological problems named above should be kept in mind. We will only deal with the question of form-meaning relationships here.

With respect to the formal appearance of allomorphs we have claimed that a large amount of allomorphy is more complex than a small amount. As validation from second language learners shows, plural formation of languages with a high number of allomorphs – such as German – is more difficult to acquire (Jordens 2004: 1808). Learners tend to use a reduced set of allomorphs for regularized plural formation (Wegener 1994: 277). We have also claimed that fused encoding of different functions is more complex than transparent encoding. Validation shows that this is mapped in the order of first language acquisition by children: "A morpheme marking x is acquired before one that marks $x + y$, and so on" (Clark 1998: 377).

Concerning the formal techniques, we have claimed that affixation is the simplest whereas stem involvement is fairly complex. Generalized, validation implies that this also corresponds with user difficulties: second language learners avoid umlaut plurals in German (Wegener 1994). Contrary to our expectations, zero morphology does not seem to cause complications to second language learners of German, which can be a hint that validation does not correspond to the theoretical considerations on complexity that we have examined. This shows that validation can also contradict our complexity metric, but further validation data needs to be accounted in order to reach this conclusion.

These examples can only offer a small impression of the results we hope to achieve in the validation step. Of course, when one takes different user groups into account, the results will not be as clear-cut as in the examples. The exemplary studies imply that for some specific cases the results of our analysis of morphological complexity correlate with difficulties for language users. Only if this is the case in further, more detailed and more general validation studies – keeping in mind the methodological problems outlined above – can the linguistic relevance of complexity metrics be confirmed.

In conclusion, in this article we have proposed a metric for analyzing morphological complexity by means of multidimensional criteria. We have examined complexity as a system property and as a theoretical issue modelled with respect to an ideal language user, and we have based our conclusions on contrastive insights from the comparison of ten Germanic languages. According to the multi-dimensional criteria the languages were ranked on a complexity scale. We could avoid problems of comparability of the established criteria by giving them equal weight and ranking the languages depending on their diffusion. The metric's relevance remains to be investigated in a validation step, which was outlined as an area of future research.

Abbreviations

ACC	accusative
DAT	dative
NOM	nominative
PL	plural

References

Allan, R., Holmes, P. & Lundskær-Nielsen, T. 1998. *Danish*: *A Comprehensive Grammar*. 2nd ed. London and New York: Routledge.

Augst, G. 1975. Zum Pluralsystem. In *Untersuchungen zum Morpheminventar der deutschen Gegenwartsprache*, G. Augst, 5–70. Tübingen: Narr.

Booij, G.E. 1996. Inherent versus contextual inflection and the split morphology hypothesis. *Yearbook of morphology* 1995: 1–16.

Bybee, J.L. 1985. *Morphology*. *A Study of the Relation between Meaning and Form*. Amsterdam: John Benjamins.

Bybee, J.L. 1994. Morphological universals and change. In *The Encyclopedia of Language and Linguistics*, vol. 5, R.E. Asher (ed.), 2557–2562. Oxford: Pergamon.

Christophory, J. 2004. *Mir schwätze lëtzebuergesch*. Luxembourg: Éditions Paul Bauler.

Clark, E.V. 1998. Morphology in language acquisition. In *The handbook of morphology*, A. Spencer & A.M. Zwicky (eds), 374–389. Oxford: Blackwell.

Dahl, Ö. 2004. *The Growth and Maintenance of Linguistic Complexity* [Studies in Language Companion Series 71]. Amsterdam: John Benjamins.

Dammel, A., Kürschner, S. & Nübling, D. (manuscript). Plural allomorphy in ten Germanic languages – assignment and exponence in diachrony.

Derrmann-Loutsch, L. 2004. *Deutsch-Luxemburgisches Wörterbuch*. 2nd ed. Luxembourg: Éditions Saint Paul.

Donaldson, B. 1994. Afrikaans. In *The Germanic languages*, E. König & J. van der Auwera (eds), 478–505. London, New York: Routledge.

Dressler, W., Mayerthaler, W., Panagl, O. & Wurzel, W.U. 1987. *Leitmotifs in Natural Morphology* [Studies in Language Companion Series 10]. Amsterdam: John Benjamins.

Duden. Die Grammatik. 2005. 7th ed. [Duden 4]. Mannheim [etc.]: Dudenverlag.

Eisenberg, P. 2004. *Grundriß der deutschen Grammatik, vol. 1: Das Wort*. Stuttgart [etc.]: Metzler.

Hawkins, J.A. 2004. *Efficiency and Complexity in Grammars*. Oxford/New York: Oxford University Press.

Haeseryn, W., Romijn, K., Geerts, G., de Rooij, J. & van den Toorn, M.C. 1997. *Algemene Nederlandse Spraakkunst*, 3 vol. Groningen: Nijhoff.

Holmes, P. & Hinchliffe, I. 1994. *Swedish: A Comprehensive Grammar*. London/New York: Routledge.

Jordens, P. 2004. Second language acquisition. In *Morphology: An International Handbook on Inflection and Word-Formation*, Vol. 2 [HSK 17], G.E. Booij, C. Lehmann & J. Mugdan (eds), 1806–1816. Berlin/New York: De Gruyter.

Keller, R.E. 1961. Luxemburgish. In *German Dialects. Phonology and Morphology with Selected Texts*, R.E. Keller, 248–298. Manchester: Manchester University Press.

Köpcke, K.-M. 2000. Chaos und Ordnung – Zur semantischen Remotivierung einer Deklinationsklasse im Übergang vom Mittelhochdeutschen zum Neuhochdeutschen. In *Angemessene Strukturen: Systemorganisation in Phonologie, Morphologie und Syntax*, A. Bittner, D. Bittner & K.-M. Köpcke (eds), 107–122. Hildesheim etc.: Olms.

Kress, B. 1982. *Isländische Grammatik*. Leipzig: Enzyklopädie.

Kusters, W. 2003. *Linguistic Complexity. The Influence of Social Change on Verbal Inflection* [LOT 77]. Leiden: LOT.

Markey, T.L. 1981. *Frisian*. The Hague [etc.]: Mouton.

McWhorter, J.H. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology* 5: 125–166.

Miestamo, M. 2006. On the feasibility of complexity metrics. In *Finest linguistics. Proceedings of the Annual Finnish and Estonian Conference of Linguistics. Tallinn, May 6–7, 2004* [Publications of the Department of Estonian of Tallinn University 8], K. Kerge & M.-M. Sepper (eds) 11–26. Tallinn: TLÜ.

Neef, M. 2000a. Phonologische Konditionierung. In *Morphology: An International Handbook on Inflection and Word-Formation* [HSK 17], G.E. Booij, C. Lehmann & J. Mugdan (eds), Vol. 1, 463–473. Berlin/New York: De Gruyter.

Neef, M. 2000b. Morphologische und syntaktische Konditionierung. In *Morphology: An International Handbook on Inflection and Word-Formation* Vol. 1 [HSK 17], G.E. Booij, C. Lehmann & J. Mugdan (eds), 473–484. Berlin/New York: De Gruyter.

Nichols, J. 1992. *Linguistic Diversity in Space and Time.* Chicago/London: University of Chicago Press.

Nübling, D. 2006. Zur Entstehung und Struktur ungebändigter Allomorphie: Pluralbildungsverfahren im Luxemburgischen. In *Perspektiven einer linguistischen Luxemburgistik. Studien zu Synchronie und Diachronie*, C. Moulin & D. Nübling (eds), 107–125. Heidelberg: Winter.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. 1985. *A Comprehensive Grammar of the English Language*. Harlow: Longman.

Raidt, E.H. 1983. *Einführung in Geschichte und Struktur des Afrikaans*. Darmstadt: Wiss. Buchgesellschaft.

Tiersma, P.M. 1999. *Frisian Reference Grammar*, 2nd ed. Ljouwert: Fryske Akademy.

Thráinsson, H., Petersen, H.P., Jacobsen, J. i L. & Hansen, Z. 2004. *Faroese. An Overview and Reference Grammar*. Tórshavn: Føroya Fróðskaparfelag.

Wegener, H. 1994. Variation in the acquisition of German plural morphology by second language learners. In *How Tolerant is Universal Grammar? Essays on Language Learnability and Language Variation*, R. Tracy & E. Lattey (eds), 267–294. Tübingen: Niemeyer.

Wurzel, W.U. 1990. Morphologisierung – Komplexität – Natürlichkeit: Ein Beitrag zur Begriffsklärung. In *Spielarten der Natürlichkeit – Spielarten der Ökonomie. Beiträge zum 5. Essener Kolloquium über "Grammatikalisierung: Natürlichkeit und Systemökonomie" vom 6.10.–8.10. 1988 an der Universität Essen*, Vol. 2.1, N. Boretzky, W. Enninger & T. Stolz (eds), 129–153. Bochum: Brockmeyer.

Wurzel, W.U. 1998. On markedness. *Theoretical Linguistics* 24: 53–71.

PART III

Creoles and pidgins

The simplicity of creoles in a cross-linguistic perspective

Mikael Parkvall
Stockholms Universitet

This paper discusses the possibility of quantifying complexity in languages in general, and in creoles in particular. It argues that creoles are indeed different from non-creoles, primarily in being less complex. While this has been argued before, this is the first attempt to prove it through the use of an extensive typological database. It is noteworthy that the differing complexity is not related to the relative lack of morphology in creoles, since they are also simpler than analytical languages. Finally, the parallels between pidgins and creoles (and in particular the fact that languages sociologically intermediate between the two categories are also structurally intermediate) support the increasingly questioned belief that pidgins are born out of pidgins.

1. Introduction

A dogma in modern linguistics is that all languages are equally expressive. While they may have lexicalized different chunks of reality, they are thought to structurally possess the same expressive potential, and thus be equally suited to encode human experience. I do not intend to question this assumption. However, I do object to jumping to the conclusion that since all languages are equally expressive, they must by definition also be equally complex.

In various guises, such an assertion is found in e.g., Atchison (2001: 253), Crystal (1987: 6), Fromkin and Rodman (1974: 26), Hockett (1958), Smith (1999: 168) and countless other works.

In the following, I shall explore one possibility of ranking languages with regard to their structural complexity (or perhaps rather *one aspect* of complexity), and also address the issue of whether "the world's simplest grammars are creole grammars", as claimed by McWhorter (2001).[1]

---

1. In this paper, "complexity" is used in the same way as does McWhorter (2001). This means that an expression is more complex than another if it involves more rules, i.e., if it requires a lengthier description. No claims are made with regard to learning difficulties or other psycholinguistic aspects.

While linguists shun from claiming that a given language is more or less complex than another, they normally do not have any problem admitting that *constructions* vary with regard to complexity. A typical example would be the passive voice. A passive sentence in a language such as English contains more material than an active one, and many theoretical frameworks would have it that it is derived from its active counterpart through the application of a rule (or set of rules).

If we admit that different sentence structures vary in complexity within a single language, there is no reason to assume that this would not be possible in cross-linguistic comparisons. For instance, (1a) in Classical Chinese (from Norman 1988: 107) contains less material than its English translation in (1b).

(1a)	Classical Chinese:	*Zhì*	*zé*	*xíng*	*yǐ*
		arrive	CONJ	leave	PARTICLE
(1b)	English:	*When*	*he*	*arrive-d,*	*he*
		SUBJUNCTION	3SG.M.SUBJ	arrive-PAST	3SG.M.SUBJ
		had	*already*	*left*	
		AUX.PAST	already	leave.PAST	

Inversely, many would no doubt agree that the English translations in (2b) and (3b) are simpler than the corresponding Kathlamet Chinook (Mithun 1999: 386) or Temirgoi Adyghe (Hewitt 2005: 127) sentences:

(2a) Chinook: *ik-u-kua-qi.uit-x̣it*
IMMEDIATE.PAST-3PL.ABS-AGENTIVE-rest-PASSIVE

(2b) English: *They slept*
3PL.SUBJ sleep.PAST

(3a) Adyghe: *txəɬə-m s-j-a-d͡ʒa-n-aw*
book-DEF OBL-1SG-3SG-INDIR.OBJ-read-FUT
sə-qa-kʷʼa-ːʁ
ABS-1SG-hither-come

(3b) English: *I came to read the book*
1SG.SUBJ come.PAST COMPL read DEF book

Of course, Classical Chinese, Chinook and Adyghe are typologically rather different from English, but even related languages such as Russian and French are adorned with grammatical markings to extents differing from English:

(4a)	English:	*He*	*is*	*a*	*student*
(4b)	French:	*Il*	*est*	Ø	*étudiant*
(4c)	Russian:	*On*	Ø	Ø	*student*
		3SG.M	COP.3SG.PRES	INDEF	student

While the average linguist might agree that we are indeed dealing with differing degrees of complexity here, the standard objection is that complexity in one area of

grammar is compensated by relative simplicity in another. There is certainly some merit to this assumption – the English version of (4) is the most complex one, but we all know that French and Russian are equipped with intricacies quite unknown to English, such as rather arbitrary grammatical gender distinctions, extensive agreement, and pervasive case marking. And while Sinitic languages may look comfortably analytic (and at times even telegraphic) to a westerner, it is well known that they have, for instance, fairly complex tone systems and large sets of numeral classifiers. The communicative efficiency that the realities of human life require from languages would seem to be similar across societies, and given identical mental and articulatory equipments, it is indeed not unreasonable to expect languages all over the world to be equally complex.

However, human minds are clearly capable of handling more structure than is found in any one language. Many people grow up with two (or even more) mother tongues, and thus spend their childhood storing more linguistic data in their brains than did those of us who were raised in a monolingual environment. This appears not to pose any insurmountable obstacles to full acquisition, and a language with more complex structure than, say English, French or Russian would therefore not be incompatible with our language faculty – it would still be possible to acquire natively.

One might surmise that any excessive complexity would automatically wither away, as humans not only demand expressiveness from their languages, but also efficiency. Surely, the history of any language is replete with examples of assimilations and simplificatory reductions. But then, if all languages represent an optimal cost-benefit trade-off, all that disappears would have to somehow be replaced in order to maintain the precarious balance. To some extent this does hold true. Latin case endings were lost in its Romance offspring, but prepositions and a stricter word order compensate for it. For other features, though, it is not obvious how a loss was compensated for. As the only major European language, English has shed its former system of grammatical gender marking, and to my knowledge, no one has proposed that its job (whatever that would be!) has been taken over by any other mechanism. Indeed, in many of the case studies in Kusters (2003), it is quite obvious that reduction of a particular subsystem did *not* lead to expansion of another. And as Kusters (2003: 11) himself points out, for the conception that "complexities balance each other out" to hold, it would have to be shown that this is *always* the case, since no one denies that it sometimes is.

The very least that we must conclude, then, is that it is not scientifically responsible to *a priori* claim that all languages are equally complex. They might well be, but they need not.

2. The relationship between expressiveness and complexity

The ideological roots of the equal-complexity dogma are not particularly difficult to identify. In the early 20th century, institutional racism gradually gave way to cultural

relativism. There was a need to convince people that the languages of technologically less advanced (from a western point of view) societies were by no means primitive. A classic illustration of the new line of thought was provided by Sapir (1921: 218-219), who wrote that "When it comes to linguistic form, Plato walks with the Macedonian swineherd, Confucius with the head-hunting savage of Assam".

These days, racism has ceased to be a salient feature of academic discourse. Yet, the concept that all languages are equally complex has become so rooted in the discipline that it is difficult to bring up for discussion, and is at best considered to be outside the scope of modern linguistics (Dixon 2003: 169; Kusters 2003: 4). As Dixon (2003: 171) observes, taking the next step forward "might be confused with returning to the first stage", i.e., again embracing Eurocentrism. There is a crucial difference, however, between Dixon's "first stage" and the proposed "third stage". In the bad old days, complexity was seen as intrinsically linked to expressive power – complex was good, and simple was bad, and that was that. Indeed, when certain languages have been claimed to be simpler than certain others, this has resulted in outcries of indignation (well exemplified in DeGraff 2001a) from readers who clearly perceive the claim to be that the purportedly simpler languages are less expressive and intrinsically less worth than the allegedly more complex ones.

However, as used by Mühlhäusler (1997: 128), Posner (1997: 122) and myself, the concept of simplification has nothing to do with reduction of expressive power (for which Mühlhäusler reserves the term *impoverishment*). Indeed, given that I here treat the equal expressiveness of languages as axiomatic, a simple language could rather be seen as a more "efficient" one, in the sense that it does the same job at a lesser cost.

A hypothetical language consisting only of the schwa phoneme would certainly be both less complex and less expressive than actually existing human languages. To take a real-life example, Busnel and Classe (1976: 87, 108–109) conclude that the whistled languages they examine are indeed simpler than ordinary speech, but they also show that this is at the expense of redundancy and preciseness, with frequent misinterpretations as a result. To some extent, thus, complexity and expressive power do go hand in hand. But at the level of human languages (which, after all, are so complex that no one in the history of mankind has ever managed to produce an ideally exhaustive reference grammar), this is not necessarily so. The simplest way of demonstrating this is through a parasitic cryptolect such as *Pig Latin*. Such a variety is arguably more complex (albeit only marginally) than English in that it contains one additional rule: "take the onset of each word's first syllable and add it to the end of the word, followed by *ay*". Yet, Pig Latin does not allow the speaker to express anything that English cannot. If it were the language of the ruling class, Pig Latin would perhaps be thought of as superior to English, but structurally speaking, it is only more complex without being more expressive. Inversely, we could imagine a variety of English which is identical to today's, only without any irregularities. This fictious English would be simpler, but again, neither more nor less expressive. In a sense, Esperanto, with its exceptionless grammar is such

a language, for it basically allows a finite number of rules to be applied without (in theory) any restrictions whatsoever. While Esperanto is an artificial language (and as such often scorned by professional linguists), it has proven capable of producing literary and scientific texts, and – more importantly, given the assumption that all natively spoken languages are expressively equal – it does have native speakers (Bergen 2001; Corsetti 1996; Corsetti et al. 2004; Versteegh 1993).

For the above reasons, therefore, it is perfectly possible to claim that all languages have the same expressive potential, while not necessarily being equally complex.

3. Material and method

If, again, we take the expressivity part as axiomatic, is there any way we could investigate the degree of complexity? It is surely not possible to prove the differing-complexity hypothesis beyond any doubt, but, I claim, it is possible to adduce evidence that strongly supports it. The basic assumption behind my reasoning is simply that a complex language is a language with more complex constructions. If English has a passive, while Classical Chinese did not (Norman 1988: 101), then English is more complex *with regard to that specific construction*. This complexity may or may not be compensated for elsewhere, but if we studied a large number of constructions, and English turned out as the more complex in every single case, then, I propose, it would be justified to claim that English were more complex than Classical Chinese.

In 2005, The *World Atlas of Linguistic Structures* (Haspelmath et. al (eds) 2005; henceforth WALS) was published. The WALS database is the most extensive typological survey to date. If any material allows a large-scale assessment (for whatever purpose) of similarities and difference between languages, WALS certainly offers the best possibilities. It contains data on more than 2 500 languages, and discusses more than 140 linguistic features. The wealth is unevenly distributed, though. Almost half of all languages are represented by 15 features or less, and the ten best covered are all big languages, such as (in descending order of frequency) English, French, Finnish, Russian, Spanish, Turkish, Hungarian, Indonesian, Japanese and Mandarin. Yet, WALS includes close to 60 000 data points, which is unparalleled by any other database (at least any that is publicly available).

Among the features discussed in WALS, most are such that they do not have any obvious positive and negative value with regard to complexity. Typical representatives of this category are the word order features – it is not obvious that SOV is neither more nor less complex than SVO. Nor is it self-evident that the presence of an unusual feature (such as click phonemes or object-initial word order) is more complex than a common one in its place. Cross-linguistic frequency is a matter of statistics, and within a given system, fixed stress on the final syllable does not require a more complex specification than stressing of the initial one (which is more common). In many

cases, languages simply do things differently, without this having any bearing on the issue of complexity.[2]

A third category of features which I have made an effort to exclude is synthesis. A high degree of synthesis is certainly perceived as being complex by laymen, and so far as "gut feeling" is concerned, also by many of us linguists. It is certainly true that synthesis tends to *lead to* arguably complex structures, such as fusion. Yet, synthesis *in itself* is not complex by my use of the term. Danish *bogen* 'the book' (where *-en* is the non-neuter definite article) differs from Spanish *el libro* only by being synthetic, but otherwise requires the same grammatical machinery to produce. There is also a pedagogical reason to exclude synthesis – we already know that the languages of the world differ considerably with regard to this, and a complexity measure including it might do little but confirm the impression that puts Greenlandic and Vietnamese at opposite ends of the scale. That said, the material used here does not allow to circumnavigate the synthesis issue completely – for some of the features, WALS only indicates the presence of a bound form versus the complete absence. This, however, applies to a small minority of features, as indicated in the list below, and its effect seems marginal, as we shall see.

In short, while any feature could be subject to debate, my aim is to use those that are likely to cause the least controversy among linguists. For all the traits listed, I consider their presence (or the presence in larger numbers) to add to the overall complexity of a language.

In some cases, as discussed earlier, we could expect an absent feature to be compensated for by the presence of something not studied by WALS, but I have made some effort to exclude such traits.[3] Of course, a language not having a given feature is normally *able* to express it, but it does not *have to*, and may not even have a conventionalized strategy to unambiguously do so. It is at the very least difficult to identify what in Standard Average European would correspond to the duals, the several degrees of "pastness", the antipassives, the two dozen noun classes, the numeral classifiers, the inclusivity, evidentiality and alienability distinctions and the distinction between various copula types found in certain more "exotic" languages. Again, while a language such as English is able to indicate the distinctions made by evidentiality affixes in Amazonia by circumlocutions (*It is said that. . .*, etc.), these are by no means obligatory, and do not

2. Sometimes, crosslinguistic frequency and complexity are even opposed to one another: the vast majority of languages require their speakers to make a distance contrast in demonstratives (WALS, Map 42), while a demonstrative system without this elaboration (everything else being equal) would clearly require a shorter description.

3. A typical example is that of case affix inventories, for which WALS provides data. There is none, however, on the size of the adposition arsenal, whose members in case-less languages tend to correspond to (at least locative) cases.

form a central part of the grammar of English. Evidentiality in English is purely a matter of pragmatics, rather than distinctions imposed on the speaker by her language.

Importantly, if the correspondences between a feature in language A and a compensatory feature in language B cannot be made, the only reason to assume that it is at all compensated for is the idea that differences *must* somehow balance each other out – and this clearly makes the reasoning circular. In cases such as that of evidentiality, we can at least find a periphrastic correspondence in English, but in the case of, say, gender, there is not even a way of translating the gender tag that every NP in certain other languages carries with it. French *la table* consists of a definite article, the noun 'table' and a gender marker, of which only the first two correspond to anything whatsoever in English. There is simply no way the gender marker could be expressed in English without making reference to French (especially not since gender is language-specific, with 'table' being masculine in e.g., German and Russian).

It could be argued (and rightly so) that such distinctions make a contribution to the overall redundancy of a language, in which case gender might as well correspond to, say, a larger phoneme inventory or longer morphemes. But then again, this presupposes that all languages have a similar (and constant!) degree of redundancy, which again makes the argument circular.

Having set the basis of my reasoning, here follow the data from WALS chosen for further consideration:

Table 1. WALS features included in the study.

	Feature[a]	**WALS map no.**
F01	Size of consonant inventories	1
F02	Size of vowel quality inventories	2
F03	Phonemic vowel nasalization	10
F04	Complexity of syllable structure	12
F05	Tone	13
F06	Overt marking of direct object	23
F07	Double marking of direct object	23
F08	Possession by double marking	24
F09	Overt possession marking	24
F10	Reduplication	27
F11	Gender	30–32
F12	Number of genders	30
F13	Non-semantic gender assignment	32
F14	Grammaticalized nominal plural	33–34
F15	Definite articles	37
F16	Indefinite articles	38
F17	Inclusivity (in either pronouns or verb morphology)	39–40
F18	Distance contrast in demonstratives	41

(*Continued*)

Table 1. Continued.

	Feature[a]	WALS map no.
F19	Gender in pronouns	44
F20	Politeness in pronouns	45
F21	Person marking on adpositions	48
F22	Comitative ≠ instrumental	52
F23	Ordinals exist as a separate class beyond 'first'	53
F24	Suppletive ordinals beyond 'first'	53
F25	Obligatory numeral classifiers	55
F26	Possessive classification	59
F27	Conjunction 'and' ≠ adposition 'with'	63
F28	Difference between nominal and verbal conjunction	64
F29	Grammaticalized perfective/imperfective	65
F30	Grammaticalized past/non-past	66
F31	Remoteness distinctions of past	66
F32	Morphological future	67
F33	Grammaticalized perfect	68
F34	Morphological imperative	70
F35	Morphological optative	73
F36	Grammaticalized evidentiality distinctions	76
F37	Both indirect and direct evidentials	77
F38	Non-neutral marking of full NPs	98
F39	Non-neutral marking of pronouns	99
F40	Subject marking as both free word and agreement	101–102
F41	Passive	107
F42	Antipassive	108
F43	Applicative	109
F44	Obligatorily double negation	112
F45	Asymetric negation[b]	113
F46	Equative copula ≠ Locative copula	119
F47	Obligatorily overt equative copula	120
F48	Demonstratives marked for number	–
F49	Demonstratives marked for gender	–
F50	Demonstratives marked for case	–
F51	Total amount of verbal suppletion	–
F52	Alienability distinctions	–
F53	Number of pronominal numbers	–

[a] It is not possible to provide detailed descriptions of each feature here, so for precise definitions, please refer to WALS.

[b] This basically means that the negated sentence differs from its positive counterpart in other ways than the mere addition of a negator morpheme.

A few features not present in the WALS database were added. F48-50 were taken from Diessel (1999: 171–173), while F51 was provided by Ljuba Veselinova (p.c.) – these data are the "leftovers", as it were, from her contributions to WALS and from Veselinova (2003). F52 was taken from Nichols (1992: 294-301), and F53 from Harley and Ritter (2002). The additions of these (but not other) data were done simply because I happened to have access to them. With the exception of Harley and Ritter, all the authors concerned are also contributors to WALS, and the added material largely consists of elaborations of their respective WALS articles.

In some cases, we could wish for more data that might reveal if the absence of a feature is indeed compensated for elsewhere in the grammar. However, in one sense, there is a point in using a predefined data source. WALS is produced by the world's most renowned typologists, whose combined expertise is second to none. They have chosen their features purely out of an interest in typology, and without the hidden agenda that I could, would, and have been accused of having.

Most features have binary YES/NO values (occasionally with an intermediate value) which easily translate into the numbers 1 and 0 (or 0.5, as the case may be). Those which do not are F01, F02, F04, F05, F12, F18, F20, F26, F51 and F53. Some of these are numerically coded, and in those cases I compressed the values to fit the zero-to-one range. Thus, the pronominal numbers (F53), which range from zero to four, were straightforwardly converted into the format 0, 0.25, 0.5, 0.75 and 1. In the remaining cases, the values were expressed in prose in WALS. Syllable structures (F04), for instance, were classified in the source as "simple", "moderately complex" and "complex". In my grid, these classifications were converted into 0, 0.5 and 1. For all features, thus, the possible range is from zero (minimally complex) to one (maximally complex).[4]

As mentioned earlier, there are plenty of gaps in the WALS database. I therefore decided to restrict myself to the languages with values for at least 30 out of 53 features considered. This results in 155 languages and a total coverage above 80% (i.e., blank cells amounting to less than 20% in a 155 × 53 grid). The total score for each language is divided by the number of features included for that language, and thus, Ladakhi (with 31 attestations) stands a chance to get an average score on par with Basque, for which all 53 cells are filled in.

4. As pointed out by Matti Miestamo (p.c.), the use of 0.5 for intermediate values has its problems. For instance a language that has both symmetric and asymmetric negation (F45) may have much more complex asymmetry than a language where negation is always asymmetric. While this is clearly something that requires future revision, I am convinced that the overall effect of this on the following calculations is minimal.

4. Results

The results of the calculations are tabulated in Table 2 below:

Table 2. Complexity ranking of WALS languages.

Rank	Language	Complexity score	Rank	Language	Complexity score
1	Burushaski	0.62	79	Fijian	0.41
2	Copainalá Zoque	0.57	80	Amele	0.41
3	Khoekhoe	0.57	81	Khalkha	0.41
4	Beja	0.56	82	Oneida	0.41
5	Koasati	0.55	83	Chamorro	0.41
6	Kannada	0.55	84	Maung	0.41
7	Ladakhi	0.54	85	Malagasy	0.40
8	Abkhaz	0.54	86	Kewa	0.40
9	Hindi	0.54	87	Ainu	0.39
10	Latvian	0.53	88	Mezquital Otomí	0.39
11	Spanish	0.53	89	Mandarin	0.39
12	Greek	0.53	90	Arapesh	0.39
13	Grebo	0.52	91	Plains Cree	0.39
14	Basque	0.52	92	Semelai	0.39
15	Ingush	0.52	93	Nez Perce	0.39
16	French	0.51	94	Egyptian Arabic	0.39
17	Kanuri	0.51	95	Ju|'hoan	0.39
18	Supyire	0.50	96	Nivkh	0.39
19	Maricopa	0.50	97	Hungarian	0.38
20	Passamaquoddy-Maliseet	0.50	98	Maori	0.38
21	Evenki	0.50	99	Kayardild	0.38
22	Comanche	0.50	100	Igbo	0.38
23	Yaqui	0.50	101	Yurok	0.38
24	Mundari	0.49	102	Urubú-Kaapor	0.38
25	Alamblak	0.49	103	Burmese	0.38
26	Dyirbal	0.49	104	Zuni	0.37
27	German	0.49	105	Acoma	0.37
28	Meithei	0.49	106	Asmat	0.37
29	Korean	0.49	107	Ngiyambaa	0.37
30	Slave	0.49	108	Imbabura Quechua	0.37
31	Lezgian	0.49	109	Wari'	0.37
32	Georgian	0.49	110	Yoruba	0.37
33	Jakaltek	0.49	111	Bagirmi	0.37
34	Hunzib	0.48	112	Kolyma Yukaghir	0.36
35	Zulu	0.48	113	Martuthunira	0.36
36	Karok	0.48	114	Gooniyandi	0.36

(*Continued*)

Table 2. Continued.

Rank	Language	Complexity score	Rank	Language	Complexity score
37	Hebrew	0.48	115	Rama	0.36
38	Paiwan	0.48	116	Iraqw	0.36
39	Nenets	0.48	117	Suena	0.35
40	Tagalog	0.47	118	Chukchi	0.35
41	Swedish	0.47	119	Ket	0.35
42	Kunama	0.47	120	Brahui	0.35
43	Mangarrayi	0.47	121	Lango	0.34
44	Diola-Fogny	0.47	122	Yagua	0.34
45	Luvale	0.47	123	Awa Pit	0.33
46	Japanese	0.46	124	Tiwi	0.33
47	Middle Atlas Berber	0.46	125	Canela-Krahô	0.33
48	Eastern Armenian	0.46	126	Rapanui	0.33
49	Swahili	0.46	127	Ewe	0.32
50	Persian	0.46	128	Haida	0.32
51	Lower Grand Valley Dani	0.46	129	Koyraboro Senni	0.32
52	Hausa	0.46	130	Khmer	0.32
53	Krongo	0.46	131	Wichita	0.31
54	Kiowa	0.46	132	Thai	0.30
55	Finnish	0.45	133	Taba	0.30
56	Barasano	0.45	134	Kilivila	0.30
57	West Greenlandic	0.45	135	Ika	0.30
58	Coos	0.45	136	Tetelcingo Nahuatl	0.30
59	Turkish	0.45	137	Warao	0.30
60	Wardaman	0.45	138	Mapudungun	0.30
61	Harar Oromo	0.44	139	Cayuvava	0.29
62	Cahuilla	0.44	140	Kutenai	0.29
63	Khasi	0.44	141	Shipibo-Konibo	0.28
64	Lahu	0.44	142	Tukang Besi	0.28
65	Guaraní	0.44	143	Yidiny	0.27
66	Lavukaleve	0.44	144	Apurinã	0.27
67	Imonda	0.44	145	Chalcatongo Mixtec	0.27
68	Epena Pedee	0.44	146	Indonesian	0.26
69	Lakhota	0.43	147	Vietnamese	0.26
70	Nunggubuyu	0.43	148	Daga	0.25
71	Sanuma	0.43	149	Usan	0.23
72	Hixkaryana	0.43	150	Ndyuka	0.22
73	Russian	0.43	151	Maybrat	0.22
74	Wichí	0.43	152	Kobon	0.20
75	Southern Sierra Miwok	0.42	153	Hmong Njua	0.20
76	Yimas	0.42	154	Pirahã	0.18
77	Paumarí	0.42	155	Sango	0.15
78	English	0.42			

A striking – and, to me, unexpected – characteristic of these results is that the Indo-European languages are so well-represented at the top of the list. I am uncertain why that is so. Either this is a fair description of the state of affairs, or perhaps it illustrates subconscious Eurocentrism on behalf of me or the WALS contributors (or both).

With this exception, however, the spontaneous reaction from most of those with some grasp of typology to whom I have shown the list is that it is largely compatible with their intuitions.

5. The special case of creoles

Disregarding the position of European languages, I find the bottom ranks to be the most interesting part of the list. Positions 150 and 155 are occupied by the only two creoles in the sample – Ndyuka and Sango. Creoles have frequently been suggested to have simpler grammars than other languages, most recently by McWhorter (2001). The suggestion has been met with scepticism by many linguists (see for instance the reactions in *Linguistic Typology* 5 (2–3), 2001), and within creolistics itself by the vast majority. Nevertheless, the subject has been considered so important that even the most prestigious journal of the discipline, *Language*, took the rare step of publishing an opinion piece on the matter (DeGraff 2003). Without a single linguistic example (let alone comparative evidence), that paper dismisses ideas about the relative simplicity of creoles as 'myths', 'prejudices', 'illusions', and 'fantasies'.

In this light, it is indeed striking that the two creoles here rank among the least complex languages, both surpassed by 98% of the sample. Nevertheless, four languages turn out to be simpler than at least one of the creoles. It thus seems that, at least in the present metric, individual non-creoles can indeed be less complex than individual creoles. These four, however, do not form a class (areal, genetic or otherwise), a fact to which we shall return shortly.

6. What are creoles?

Before continuing, a few background details are necessary.

During the decades that creolistics has existed as a semi-autonomous subdiscipline, opinions on what creoles *really* are have varied considerably. Two traditional views are perhaps better known to the general linguist than others. First, some creolists have seen creoles more or less as relexifications of their substrates, i.e., old (typically African or Melanesian) grammars equipped with new (typically European) lexica. Representatives of this school are e.g., Keesing (1988) and Lefebvre (1998). Then, according to Derek Bickerton (1981, 1984), creoles offer a unique window on the human language faculty by representing overt manifestations of the unmarked parameter settings. Known as the *Language Bioprogram Hypothesis*, this view had an impact outside of

creolistics, and is occasionally still quoted as representing the creolistic state of the art. Within the subdiscipline itself, the reception was less cordial, and the Bioprogram is now dead as the dodo.

Meanwhile, the dominant tendency among Francophone creolists (prime examples being Chaudenson 1979, 1992, 1995) was to consider creoles direct descendants of their lexifiers. The relationship between, say, Haitian and French was claimed to be largely similar to that between Latin and French.

After having been viewed with suspicion by outsiders, the 1990s saw this view make advances among creolists writing in English. Propagated by in particular Salikoko Mufwene and Michel DeGraff, it has now gained a considerable following, and may perhaps represent the closest thing to a consensus that there is in creole studies. Following DeGraff (2003), I shall refer to this school of thought as *uniformitarian*.

Two of the most central claims made by the uniformitarian school are 1) that creoles do not derive from pidgins, but instead fit equally well into a family tree model as other languages do (Chaudenson 1995: 66; Mufwene 2000, 2001: 152, 153; DeGraff 2002: 377, 378, 2003: 398, 399), and 2) that the very label "creole" has no structural/typological correlate, and is solely defined by its history (Chaudenson 2001: 145; DeGraff 2003: 391; Mufwene 2001: 10, 2002, 2003: 277–278). A language is/should be/can be classified as a creole exclusively on the basis of its past. Mufwene (2000) defines creoles as "a group of vernaculars whose developments are similar especially in their temporal and geographical positions, viz., in tropical colonies settled by Europeans practicing slave-based economy from the 17th to the 19th centuries. The lexifiers are typically non-standard varieties of European languages" (similar definitions are found in Mufwene 2003 and DeGraff 2003: 391).

The results from the present study do not confirm the view on creoles supported by uniformitarians.

7. Adding more contact languages

The presence of only two creoles in the sample (Ndyuka and Sango) is somewhat problematic, however, so I entered as much information as I could from another 30 pidgins and creoles, resulting in the following values.[5]

5. Reference grammars are available for only some of these languages, so much of the data were taken from sources too numerous to mention here. The most important sources, however, are Bailey (1966), Baker (1972), Barrena (1957), Baxter (1988), Bollée (1977), Broch and Jahr (1984), Carrington (1984), Corne (1977), Damoiseau (1984), Ehrhart (1993), Ferraz (1979), Günther (1973), Heine (1982), Kihm (1994), Kouwenberg and Murray (1994), Munteanu (1996), Scantamburlo (1981) and Verhaar (1995) (details available on request). In cases where several sociolects have been identified, I am referring here to the basilect, i.e., the variety the furthest removed from the lexifier. The same 30-feature limit as used before applies here as well.

Table 3. Complexity values of pidgins and creoles added to the sample.

Language	Main lexifier	Area	Type[a)]	Score
Annobonese	Portuguese	West Africa	creole	0.22
Australian Creole	English	Pacific	creole	0.16
Bislama	English	Pacific	expanded pidgin	0.13
Chinook Jargon	Chinook	North America	(expanded) pidgin	0.15
Dominican	French	Caribbean	creole	0.21
Fanakalo	Zulu	South Africa	pidgin	0.20
Guadeloupean	French	Caribbean	creole	0.25
Guinea Bissau Creole	Portuguese	West Africa	creole	0.27
Haitian	French	Caribbean	creole	0.29
Jamaican	English	Caribbean	creole	0.27
Kinubi	Arabic	East Africa	creole	0.24
Krio	English	West Africa	creole	0.28
Lingua Franca	Italian/Romance	Mediterranian	pidgin	0.06
Martinican	French	Caribbean	creole	0.20
Mauritian	French	Indian Ocean	creole	0.20
Negerhollands	Dutch	Caribbean	creole	0.20
Nigerian Pidgin	English	West Africa	expanded pidgin	0.25
Palenquero	Spanish	South America	creole	0.27
Papia Kristang	Portuguese	South-East Asia	creole	0.18
Papiamentu	Spanish	Caribbean	creole	0.33
Principense	Portuguese	West Africa	creole	0.23
Russenorsk	Norwegian & Russian	Arctic Ocean	pidgin	0.00
Sãotomense	Portuguese	West Africa	creole	0.31
Saramaccan	English	South America	creole	0.25
Seychellois	French	Indian Ocean	creole	0.25
Sranan	English	South America	creole	0.24
St. Lucian	French	Caribbean	creole	0.22
Tayo	French	Pacific	creole	0.22
Tok Pisin	English	Pacific	expanded pidgin	0.22

[a)] The labels "pidgin" and "creole" are used here in the traditional way, i.e., for non-native and native varieties respectively. "Expanded pidgin" refers to a variety which is mainly used as an L2, but which serves as the primary (and in some cases even native) language for some of its users.

As with the sample already discussed, there are gaps in the additional creole data, and it would be surprising if there were no errors or debatable classifications among these almost 1 200 additional data points. In case the suspicion should arise that these additional values were tailor-made to suit my preconceptions, it should be noted that these additional languages score slightly *higher* than Sango and Ndyuka, the two creoles for which data were provided by WALS. While the addition did not so much alter

the position of creoles, it does indicate that Sango and Ndyuka are rather representative in this respect.

8. Creoles and other groups compared

Now, recall that the uniformitarian school denies that creoles constitute a typological class, or that they are in any way identifiable on structural grounds. Therefore, any resemblances that would make them *look like* a group would be due to chance. It should thus be possible to come up with other groupings that are superficially equally tenable.

I experimented with 90 different ways of lumping languages together. First, on the basis of geography. Secondly, I grouped languages according to family. Third, on typological grounds, using 26 randomly selected WALS features not already involved in my calculations. Fourth, 16 groups were formed on the basis of a number of sociolinguistic factors. These four groups are referred to in the table below as GEO, GEN, TYPO and SOC respectively.

It could be suspected that these factors would have some bearing on complexity, e.g., because of areal pressure, through inheritance of complexity (or simplicity), by means of factors such as typological "harmony" or sociological correlations. Finally, however, I tested a number of *intentionally nonsensical* features that are not normally believed to have anything to do with linguistic structure (indicated as SILLY in the table below).

The point here is this: if there are individual languages which are less complex than individual creoles, are there also groups of languages which are less complex than creoles? If not, such a result would further support the notion of creoles as a typological class.

The following table demonstrates the clustering of various groupings with regard to the degree of complexity:[6]

Table 4. Complexity values of various types of groupings of the WALS languages.

	Complexity score 0.30–0.39	Complexity score 0.40–0.49
GEO	Languages of Oceania, of South America, of South-East Asia, of the southern hemisphere	Languages of Africa, of Asia, of Europe, of North America, of the eastern hemisphere, of the northern hemisphere, of the western hemisphere
GEN	Austro-Asiatic, Austronesian, Pama-Nyungan, Trans-New Guinea	Afro-Asiatic, Algic, Gunwingguan, Indo-European, Isolates, Niger-Congo, Nilo-Saharan, Penutian, Sino-Tibetan, Uralic, Uto-Aztecan lgs

(*Continued*)

6. Groups with less than three representatives in the sample are excluded.

Table 4. Continued.

	Complexity score 0.30–0.39	Complexity score 0.40–0.49
TYPO[a]	Head marking languages, SVO languages, predominantly prefixing languages, languages encoding ditranstitives as double-object constructions, with 'exceed'-comparatives, without laterals, without morphological case marking	Dependent marking languages, lgs differentiating 'hand' and 'finger', in which "adjectives" are morpho-syntactically "noun-like", in which "adjectives" are morphosyntactically "verb-like", not differentiating 'hand' and 'finger', requiring overt subject pronoun, with /ŋ/, with decimal numeral systems, with glottalized consonants, with laterals, with less than average degree of synthesis, with more than average degree of synthesis, with morphological case marking, with non-decimal numeral systems, without /ŋ/, without 'exceed'-comparatives, predominantly suffixing lgs, pro-drop lgs, SOV lgs
SOC[b]	Languages spoken mainly in virtually monolingual nations, spoken mainly in countries where Spanish is official	Languages spoken in former European colonies, without official status in any country, with less than 1000 speakers, with official status in more than one country, with official status in at least on e country, with more than 1M speakers, spoken in less literate (=below world average) societies, spoken in highly literate (=above average) societies, spoken mainly in Commonwealth countries, spoken mainly in *Francophonie* countries, spoken mainly in very multilingual nations, majority languages (in their respective countries), minority languages (ditto)
SILLY[c]	Languages using Roman script, spoken by yellow peoples, by red peoples, mainly in countries whose name begin with the letter C, mainly in countries whose name begin with the letter P	Languages mentioned in Chomsky (1966), lgs spoken by black peoples, by white peoples, by Christians, by Moslems, in areas free of malaria, in malaria-infested areas, in countries with an above average mobile phone ownership, in countries with a below average mobile phone ownership, in democracies, in dictatorships, using Arabic script, using Cyrillic script, whose names begin with the letter A, whose names begin with the letter K, whose names begin with the letter M, whose names begin with the letter S

[a] The typological information in this category was taken from the following WALS maps: 7, 8, 9, 25, 26, 49, 81, 101, 105, 118, 121, 130 and 131. The "average degree of synthesis" refers to an index (produced in the same way as the complexity index) based on boundness of tense/aspect marking, number marking and negation, person marking on verbs, two measures of inflexional morphology in general and two indices of fusion (maps 20, 21, 22, 26, 33, 69, 100, 101 and 112 in WALS).

[b] "Virtually monolingual" are those with a diversity index of 0.1 or less according to the *Ethnologue* database (15th edition). "Very multilingual" nations are those with a diversity index of at least 0.9 according to the same source (for the definition of this index, see Greenberg 1956). Numbers of speakers were also derived from the *Ethnologue*.

[c] Literacy figures and data on mobile phone ownership are from the CIA World Factbook. Democracies vs. dictatorships correspond to the labels "free" and "not free" respectively, as used by the organization Freedom House. Scripts and religions were checked with the Ethnologue. Spread of malaria was defined by the World Health Organization.

The absence of a column for values below 0.30 has a simple explanation – no group, neither sensible nor nonsensical, falls into this range. One single (genetic) group, namely North Caucasian, produced an average slightly above 0.50, and as can be seen, all groups hover around 0.40.

The average value for all non-creole (and non-pidgin) languages involved here is 0.41, and deviations from this value are relatively small. It is interesting, then, to compare this to the pidgin and creole values:

Table 5. Complexity values of the contact languages and their subgroups.

Language type	Average complexity
Creoles (excluding expanded pidgins)	0.24
Creoles (including expanded pidgins)	0.23
Sango & Ndyuka only	0.19
Expanded pidgins only	0.19
All pidgins	0.14

Given the intermediate status of the "expanded pidgins", they are here included both in the pidgin and the creole categories, as well as being given separately. Sango and Ndyuka are indicated separately because they are included in the WALS sample, and so their values cannot be due to my supplying suspect data.

A *t*-test shows that the difference between creoles (and expanded pidgins) and non-contact languages is statistically significant: a P value of < 0.001 means that there is less than a one in a thousand possibility that the difference is due to mere chance.[7]

The most interesting observation here is that it is difficult to come up with *or even to invent* a category whose members behave typologically like creoles do. If we choose – as I have done – to talk about the crucial difference between creoles and non-creoles in terms of "complexity" or not, could even be considered a matter of taste. What is paramount is that creoles are different vis-à-vis non-creoles – that alone contradicts the uniformitarian position. Of some importance with regard to the birth of creoles is that the group of languages which most closely resembles them is the category of pidgins. And among these languages, the so-called expanded pidgins are not only sociologically, but also structurally intermediate between the pidgin and the creole groups. This conforms with the idea that pidgins expand into creoles – again something that the uniformitarian school denies.

Some of the parameters in Table 4 are particularly interesting. For instance, few people – including those who are convinced about the equicomplexity among

7. The standard deviation among the creoles (0.05) is also slightly less, suggesting that the group is more homogenous with regard to structural complexity than are languages in general (0.09), something that again emphasizes that they do constitute a valid typological group.

languages – would deny that creoles are rather analytic. It has sometimes been suggested that the perceived simplicity is first and foremost a product of the analytical makeup. If nothing else, however, I have apparently succeeded in avoiding such a bias by excluding most synthesis from my complexity measure – above-average synthetic languages are hardly more complex (0.42) than the members of the below-average group (0.40). In particular, the latter do not in any way approach the creoles (0.22). It is in other words perfectly possible to be analytic, and yet be as complex as the average language – it is just that this is not what creoles do. Among the typological features, several are common in creoles (for instance SVO word order, 'exceed'-comparatives, lack of morphological case marking and "verb-like" adjectives) – but again, other languages with these features are more complex than creoles.

Given the political concerns of DeGraff (2001b: 2, 2001c: 49; Saint-Fort 2002), it is notable that "languages spoken in former European colonies", languages "without official status in any country", and "languages spoken in less literate societies" do not behave differently from others.[8] Finally, even some of the SILLY features are pertinent in this context. DeGraff has repeatedly claimed that the very idea that creoles are different from other languages is rooted in racism (since speakers of the best known creoles, including the one that he discusses, are mostly black).[9] Therefore, it is worth noting that "languages spoken by white peoples" and "languages spoken by black peoples", as well as languages spoken by humans of other pigmental endowment are about equally complex. It is creoles that stand out, not the languages of black (or red, yellow or white) peoples.

It is of course possible that an unproportionate share of a given language's complexity has been tucked away in corners of the grammar that WALS did not reach. But while this is possible, and even likely, for individual languages, things are different for entire groups. Recall that it has been claimed that the label "creole" has no typological correlate whatsoever. Should it indeed be the case that the group which turns out to be simplest in my metric is in fact more complex in other areas, then *that in itself would set creoles apart from other languages* – they would then not be "the worlds simplest languages", but instead "the languages whose complexity resides in unusual places". Whichever option you choose, it is difficult to deny that there is such a thing as a "creole typological profile".

---

8. The latter is also interesting because of Chaudenson's (e.g., 2001: 159) notion that at least some aspects of complexity are upheld because of normative pressure, something that virtually presupposes literacy.

9. Views such as those presented in this paper are said to represent "neocolonial intellectual imperialism" (DeGraff 2001c: 45) and an "anti-egalitarian approach" (DeGraff 2001c: 46–47), whereby I and others uphold "mythical hierarchies among human languages, with Creoles consistently ranked at the bottom" (DeGraff 2001c: 49). Moreover this line of thinking is not only "race-based" (DeGraff 2003: 391), but even has "race theory as cornerstone" (DeGraff 2001d).

9. Conclusion

Typologically speaking, creoles stand out from languages in general, and the most salient difference is that they present a lower structural complexity. This does not necessarily have any bearing on issues regarding psycholinguistic complexity, however, and certainly not on their expressive potential.

A comparison of the complexity found within pidgins, expanded pidgins, creoles and other languages shows that the *complexity of a language correlates with its age*. This is precisely what is predicted by a scenario (the most well-known being that set out in McWhorter 2001) in which creoles emerge through broken transmission, and where complexity accretes over time.

Abbreviations

1	first person
3	third person
ABS	absolutive
AUX	auxiliary
COMPL	complementizer
DEF	definite
FUT	future
INDIR	indirect
M	masculine
OBJ	object
OBL	oblique
PL	plural
SG	singular
SUBJ	subject

References

Aitchison, J. 2001. *Language Change: Progress or Decay?* (3rd ed.). Cambridge: Cambridge University Press.

Bailey, B. 1966. *Jamaican Creole Syntax*. Cambridge: Cambridge University Press.

Baker, P. 1972. *Kreol: A Description of Mauritian Creole*. London: C. Hurst & Co.

Barrena, N. 1957. *Gramática Annobonesa*. Madrid: Instituto de Estudos Africanos/Consejo Superior de Investigaciones Científicas.

Baxter, A. 1988. *A Grammar of Kristang (Malacca Creole Portuguese)*. Canberra: Australian National University.

Bergen, B. 2001. Nativization processes in L1 Esperanto. *Journal of Child Language* 28(3): 575–595.

Bickerton, D. 1981. *Roots of Language*. New York: Karoma Publishers.

Bickerton, D. 1984. The language bioprogram hypothesis. *The Behavioral and Brain Sciences* 7: 173–188.
Bollée, A. 1977. *Le créole français des Seychelles: Esquisse d'une grammaire, textes, vocabulaire*. Tübingen: Max Niemeyer.
Broch, I. & Jahr, E.H. 1984. *Russenorsk – et pidginspråk i Norge*. Oslo: Novus.
Busnel, R.-G. & Classe, A. 1976. *Whistled languages*. Berlin: Springer-Verlag.
Carrington, L. 1984. *Saint Lucian Creole: A Descriptive Analysis of its Phonology and Morpho-Syntax*. Hamburg: Helmut Buske Verlag.
Chaudenson, R. 1979. *Les créoles français*. Paris: Nathan.
Chaudenson, R. 1992. *Des îles, des hommes, des langues. Essai sur la créolisation linguistique et culturelle*. Paris: l'Harmattan.
Chaudenson, R. 1995. *Les créoles*. Paris: Presses Universitaires de France.
Chaudenson, R. 2001. *Creolization of Language and Culture*. London: Routledge.
Chomsky, N. 1966. *Cartesian Linguistics: A Chapter in the History of Rationalist Thought*. New York: Harper & Row.
Corne, C. 1977. *Seychelles Creole grammar: Elements for Indian Ocean Proto-Creole Reconstruction*. Tübingen: Verlag Günther Narr.
Corsetti, R. 1996. A mother tongue spoken mainly by fathers. *Language Problems and Language Planning* 20(3): 263–273.
Corsetti, R., Pinto, M.A. & Tolomeo, M. 2004. Regularizing the regular: The phenomenon of overregularization in Esperanto-speaking children. *Language Problems and Language Planning* 28(3): 261–282.
Crystal, D. 1997. *English as a Global Language*. Cambridge: Cambridge University Press.
Damoiseau, R. 1984. *Éléments de grammaire du créole martiniquais*. Fort-de-France: Hatier-Antilles.
DeGraff, M. 2001a. On the origin of Creoles: A Cartesian critique of Neo-Darwinian linguistics. *Linguistic Typology* 5(2–3): 213–310.
DeGraff, M. 2001b. The mismeasure of the Creole speaker. Paper presented at the University of Rochester, 27 April 2001.
DeGraff, M. 2001c. Morphology in Creole Genesis: Linguistics and Ideology. In *Ken Hale: A Life in Language*, M. Kenstowicz (ed.), 53–121. Cambridge: MIT Press. Pagination here refers to the MS version.
DeGraff, M. 2001d. Salikoko Mufwene – from the Congo to Chicago. *Carrier Pidgin* 29.
DeGraff, M. 2002. Relexification: A reevaluation. *Anthropological Linguistics* 44(4): 321–414.
DeGraff, M. 2003. Against Creole exceptionalism. *Language* 79(2): 391–410.
Diessel, H. 1999. *Demonstratives. Form, function and grammaticalization*. Philadelphia: John Benjamins.
Dixon, R. 2003. A program for linguistics. *Turkic Languages* 7: 157–180.
Ehrhart, S. 1993. *Le créole français de St-Louis (le tayo) en Nouvelle-Calédonie*. Hamburg: Helmut Buske.
Ferraz, L.I. 1979. *The Creole of São Tomé*. Johannesburg: Witwatersrand University Press.
Fromkin, V. & Rodman, R. 1974. *An Introduction to Language*. New York: Holt, Rinehart & Winston.
Greenberg, J. 1956. The Measurement of Linguistic Diversity. *Language* 32 (1): 109–115.
Günther, W. 1973. *Das portugiesische Kreolisch der Ihla do Príncipe*. Marburg: Marburger Studien zur Afrika- und Asienkunde.

Harley, H. & Ritter, E. 2002. A feature-geometric analysis of person and number. *Language* 78 (3), 482–526.
Haspelmath, M., Dryer, M, Gil, D. & Comrie, B. (eds). 2005. *The World Atlas of Language Structures*. Oxford: Oxford University Press.
Heine, B. 1982. *The Nubi Language of Kibera – An Arabic Creole*. Berlin: Dietrich Reimer Verlag.
Hewitt, G. 2005. North West Caucasian. *Lingua* 115(1–2): 91–145.
Hockett, C. 1958. *A Course in Modern Linguistics*. New York: Macmillan.
Keesing, R. 1988. *Melanesian Pidgin and the Oceanic Substrate*. Stanford: Stanford University Press.
Kihm, A. 1994. *Kriyol Syntax. The Portuguese-Based Creole Language of Guinea-Bissau*. Amsterdam: John Benjamins.
Kouwenberg, S. & Murray, E. 1994. *Papiamentu*. Munich and Newcastle: Lincom Europa.
Kusters, W. 2003. *Linguistic Complexity: The Influence of Social Change on Verbal Inflection*. Utrecht: LOT.
Lefebvre, C. 1998. *Creole Genesis and the Acquisition of Grammar*. Cambridge: Cambridge University Press.
McWhorter, J. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology* 5(2–3): 125–166.
Mithun, M. 1999. *The Languages of Native North America*. Cambridge: Cambridge University Press.
Mufwene, S. 2000. Creolization is a social, not a structural, process. In *Degrees of Restructuring in Creole Languages*, E. Schneider & I. Neumann-Holzschuh (eds), 65–84. Amsterdam: John Benjamins.
Mufwene, S. 2001. *Ecology of Language Evolution*.: Cambridge University Press.
Mufwene, S. 2002. What do creoles and pidgins tell us about the evolution of language? Paper presented at the Conference on Language Origins, Paris, September 26–27, 2002.
Mufwene, S. 2003. Genetic linguistics and genetic creolistics: A response to Sarah G. Thomason's "Creoles and genetic relationships". *Journal of Pidgin and Creole Languages* 18(2): 273–288.
Mühlhäusler, P. 1997. *Pidgin and Creole Linguistics. Expanded and revised edition*. London: University of Westminster Press.
Munteanu, D. 1996. *El papiamentu, lengua criolla hispánica*. Madrid: Editorial Gredos.
Nichols, J. 1992. *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press.
Norman, J. 1988. *Chinese*. Cambridge: Cambridge University Press.
Posner, R. 1997. *Linguistic Change in French*. Oxford: Clarendon Press.
Saint-Fort, H. 2002. Qu'est-ce qu'une Langue Créole? (deuxième partie). *The Haitian Times* 4(31): 2.
Sapir, E. 1921. *Language: An Introduction to the Study of Speech*. New York: Harcourt & Brace.
Scantamburlo, L. 1981. *Gramática e dicionário da língua criol da Guiné-Bissau (GCr)*. Bologna: Editrice Missionaria Italiana.
Smith, N. 1999. *Noam Chomsky: Ideas & Ideals*. Port Chester: Cambridge University Press.
Verhaar, J. 1995. *Toward a Reference Grammar of Tok Pisin*. Honolulu: University of Hawai'i Press.
Versteegh, K. 1993. Esperanto as a first language: Language acquisition with a restricted input. *Linguistics* 31-33(325): 539–555.
Veselinova, L. 2003. Suppletion in verb paradigms: Bits and pieces of a puzzle. PhD Dissertation, Department of Linguistics, Stockholm University.

Complexity in numeral systems with an investigation into pidgins and creoles

Harald Hammarström
Chalmers University of Technology

This paper defines and surveys numeral systems from languages across the world. We define the complexity of a numeral system in some detail and give examples of varying complexity from different languages. The examples are chosen to illustrate the bounds on complexity that actually occur in natural languages and to delineate tricky issues of analysis. Then we contrast the complexity in numeral systems of pidgin/creole languages versus their lexifiers and versus languages generally in the world. It turns out that pidgins/creoles have slightly less complex numeral systems than their lexifiers, but probably still more complex than the world average. However, the conclusions in this respect are limited by gaps in documentation and unsystematic knowledge of the linguistic and social history of alleged pidgin/creole languages.

1. Numerals

1.1 What are numerals?

In this paper, I define numerals as follows.[1] They are

1. *spoken*,
2. *normed expressions* that are used to denote the
3. *exact number* of objects for an
4. *open class of objects* in an
5. *open class of social situations* with
6. *the whole speech community* in question.

With the first point I mean to disregard symbol combination systems, e.g., Roman numerals, that are confined to written communication, but of course most (actually all) of our primary data come from written representations of the spoken language.

---

1. Acknowledgements: The author has benefited much from working in a project to write numeral grammars in Grammatical Framework (GF) under Aarne Ranta. Mikael Parkvall and Peter Bakker have provided invaluable help with sorting out pidgins and creoles, but neither should be held accountable for any mischaracterizations in this paper.

The second point serves to exclude expressions that also denote exact numbers, but are not the normal or neutral way to say those numbers, e.g., 'eight-times-nine-and-another-two' for the normal 'seventy-four', but also to demarcate the area where the numeral system ends, which is, when there are no normed expressions.

As for the third point, languages usually have a rich set of expressions for inexact quantities, 'a lot', 'few', 'really many', 'about fifty' (but hardly *'about fifty-one') that have relatively high frequency in discourse. These are interesting in themselves but will not be included here because of their different fuzzy nature compared to exact number expressions.

Concerning the fourth point, some languages have special counting systems for a restricted class of objects (e.g., in Wavulu (Hafford 1999) for counting coconuts). These can be quite idiosyncratic and since all languages which have exact enumeration must have a means for counting an open class of objects it is better to study that.

The reason for the fifth point, the requirement on social situations, is to take a stand on so-called body-tally systems (cf. Laycock 1975; Lean 1992). A body-tally-system may be defined as follows. Assume a sequence of body parts beginning with the fingers of one hand continuing with some points along the lower and upper arm, reaching one or more points of the head, then ending with the corresponding body-parts on the opposite arm and finally hand. A number *n* is then denoted by the *n*th body-part-term in the sequence, e.g., 'nose' or 'elbow on the other side'. Typically, body-tally systems are only used in special circumstances, such as bridal price negotiations, and in other cases one would use a different numeral system or not use exact enumeration at all. The information on the social status of the body-tally numeral systems is very incomplete; for the vast majority we do not have such information, but for those in which we do, the social situation restriction applies. Body-tallying has to be done on a physically present person, and to understand what number is referred to the process must be watched, so, for instance, body-tallying numerals would be infelicitous when it is dark. For instance, de Vries (1998) found that body-tally numerals in a Bible translation could not be understood, i.e., were often mistranslated back to Indonesian by bilingual persons. Of course, there could be some other language(s), unknown to me at present, where body-tally numerals can be used in a fully open class of social situations; such a body-tally system would, accordingly, be included in the study.

Finally, regarding the sixth point, I am not interested in numeral systems which are particular to some small subsets of the speakers of the language in question (e.g., professional mathematicians) because such systems might not respond to the conditions and needs of the majority of a society.

1.2 Why study numerals?

It is true that many languages have very small numeral inventories, that is, words up to two or three and perhaps a possibility to express exact numbers up to at most ten using these and the word for hand (Hammarström, in preparation). But in languages which do not have small numeral inventories, numeral expressions form a system whose properties can be meaningfully studied in terms of complexity.

Numerals provide a good testing bed for patterns across languages given their comparatively clear semantics and modularity. As to numeral semantics, languages may differ as to which quantificational meanings they express/lexicalize, notably in approximate numeration and whether a counted set of objects constitute a group or not, but these matters are minor compared to differences languages show e.g., in verbal tense/aspect. Likewise, although not universally, numerals tend to have uniform, clearly identifiable, syntactic behaviour within a language. Also, if two languages have exact numeration for a certain range of numbers, one expects the two to give a similar functional load to these expressions, excluding possibilities such as numbers also being used for say colours or as metaphors significantly wider in one language or the other. This appears sound also in the light of the only corpus study of numeral frequencies in a language with a small numeral system (McGregor 2004: 204), which shows that 'one' and 'two' in Gooniyandi occur with comparable frequency to 'one' and 'two' in English.

Also, lots of data are available in one form or another for numerals. It seems that numerals together with pronouns, kinship terms, body part terms, and other basic vocabulary (sun, water, etc), and perhaps "sketchy" phonological inventory, are the parts of language where there exists empirical data for a really large subset of the world's known languages. One may legitimately ask just how large this subset is when it comes to numerals – for how many languages do we have data on numerals? Let us say we count about 7000 attested native spoken languages for the world. A definite lower bound is 2500, since I can produce a list of references to numeral data from 2500 definitely distinct languages. An upper bound is harder to give. I entertain the rather time-consuming methodology of trying to obtain every first-hand descriptive data reference found in any handbook or relevant publication whatsoever. I currently have about 5000 such items, some describing numeral systems of many languages in the same publication, but it is impossible to say at this point how many languages they account for since they include dialectal varieties, varieties from the same location but different centuries, partial data, data of varying quality, duplicated data, etc. I also have about a 1000 more references that I have not yet been able to obtain (which may contain further references).

2. Complexity

The following characterization and prevalence of complexity in numeral systems is based on inspection of the above mentioned set of data.

Basically I subscribe to the idea of measuring the complexity of a numeral system by the amount of information necessary to describe the forms. This clearly depends on the scope and flexibility of the means of description (the *description language*). From a computer science perspective a maximally expressive description language is a Turing machine, and it turns out that the minimum-length-description of an object can be meaningfully defined, which, up to a constant factor, is only a property of the object itself and not restricted by the poverty of the description language. The size of this minimal description of an arbitrary object, represented as a string of binary symbols, is called its *Kolmogorov*

Complexity (Vitányi and Li 1997). However, there are several reasons why we will not use Kolmogorov complexity here; in general, Kolmogorov complexity is not computable, that is, it can be proven that there is no one algorithm that will find the minimum-description for *all* possible objects (this does not contradict it being well-defined). Further, Kolmogorov complexity only gives valuable insights on asymptotic behaviour, which is not really relevant for particular finite natural language numeral systems and the level of formality required is completely foreign to traditional linguistic analysis.

Instead, I will look at complexity, i.e., minimum description length, on a level more familiar to computational linguistic analysis, namely where the description language is any standard phrase-structure-based grammar formalism and the objects to be described are strings of phonemes. I will freely ignore (morpho-)phonological alterations and other properties of the numeral forms that belong to the language as a whole rather than the numerals in particular. I will only sketch structural properties to hint how they must be described, and no exact computations of complexity will be given. The reason for this laxness is that I will really only be interested in relative complexity, e.g., if such and such creole language is less complex than such and such non-creole, so it does not really matter how I treat the details as long as I vow to treat them uniformly, but in an unspecified way.

2.1 Complexity as irregularity

We will begin with the following vague formulation of irregularity and make it more precise with examples:

> The form of a numeral is not systematically predictable from the forms of its mathematical parts and knowledge of the rest of the language.

For example, in Russian (Comrie 1992) shown in Table 1, 40 is irregular.

Table 1. The forms for the tens in Russian

10	ДЕСЯТЬ	desyat’	60	ШЕСТЬДЕСЯТЬ	šest’desyat’
20	ДВАДЦАТЬ	dvadcat’	70	СЕМЬДЕСЯТЬ	sem’desyat’
30	ТРИДЦАТЬ	tridcat’	80	ВОСЕМЬДЕСЯТЬ	vosem’desyat’
40	СОРОК	sorok	90	ДЕВЯНОСТО	devyanosto
50	ПЯТЬДЕСЯТЬ	pyat’desyat’	100	СТО	sto

Many languages have irregularities, specific to their numeral forms, that can be captured by a (sub-)rule. That is, we have several irregular forms that can be captured by one rule that is not subsumed by some greater rule. For example, many modern Indo-Aryan languages form all of 19, 29, . . ., 89 by subtraction as 1–20, 1–30, . . ., 1–90 rather than the usual additive pattern for the other numbers 11–99 (Berger 1992). Another example, Camus (a Maa Dialect in Kenya) (Heine 1980: 110–111) in Table 2, has idiosyncratic forms for the tens up to 50, after which formation can be captured

by a rule. In general, I will count complexity as proportional to the number of rules necessary, where one exception counts as much as one new rule.

Table 2. Formation of tens in Camus

1	-bô	10	tomon
2	-aré	20	tíkítam
3	-uní	30	osom
4	-oŋwán	40	ártam
5	ímíét	50	onom
6	Ile	60	n-tomon-i ile
7	sapa	70	n-tomon-i sapa
8	isiet	80	n-tomon-i isiet
9	saal	90	n-tomon-i saal

It follows that the most complex languages are those that have the largest number of irregularities. For example, in Panjabi (Shackle 2003: 602), as shown in Table 3, the numbers 1–99 (although easily etymologizable) show no consistent pattern in formation and have to be learned more or less by rote.

Table 3. Panjabi words 1–100

1	Ik	21	Ikki	41	Iktaḷi	61	Ikáṭh	81	Ikasi
2	do	22	bai	42	bətaḷi	62	báṭh	82	bIasi
3	tIn	23	tei	43	tətaḷi	63	təréṭh	83	tIrasi
4	car	24	cə́vvi	44	cUtaḷi	64	cə̃́ṭh	84	cUrasi
5	pəñj	25	pə́ñji	45	pəñjtaḷi	65	pɛ̃́ṭh	85	pəñjasi
6	che	26	chəbbi	46	chIaḷi	66	chIáṭh	86	chIasi
7	sətt	27	sətai	47	səntaḷi	67	sətáṭh	87	sətasi
8	əṭṭh	28	əṭhai	48	əṭhtaḷi	68	əṭháṭh	88	əṭhasi
9	nɔ	29	Unətti	49	Unəñja	69	Unəttər	89	Unanvẽ
10	dəs	30	tí	50	pə́ñjá	70	səttər	90	nabbe
11	gIaraã	31	Ikətti	51	Ikvəñjá	71	Ikə̀ttər	91	Ikanvẽ
12	barã	32	bətti	52	bəvəñja	72	Bə̀ttər	92	banvẽ
13	terã	33	tẽti	53	tərvəñja	73	tIə̀ttər	93	tIranvẽ
14	cədã	34	cə̃ti	54	cUrvəñja	74	cUə̀ttər	94	cUranvẽ
15	pəndrã	35	pẽti	55	pəcvəñja	75	pəñjə̀ttər	95	pəcanvẽ
16	soḷã	36	chətti	56	chIvəñja	76	chIə̀ttər	96	chIanvẽ
17	sətarã	37	sẽti	57	sətvəñja	77	sətə̀ttər	97	sətanvẽ
18	əṭharã	38	əṭhətti	58	əṭhvəñja	78	əṭhə̀ttər	98	əṭhanvẽ
19	Unni	39	Untaḷi	59	Unáṭh	79	Unasi	99	nəṛInvẽ
20	ví	40	caḷi	60	səṭṭh	80	əssi	100	sɔ

It is readily seen that the regularity of a numeral system is strongly connected to the concept of base. The set of bases of a natural language numeral system may be defined as follows.

The number *n* is a base iff

1. the next higher base (or the end of the normed expressions) is a multiple of *n*; and
2. a proper majority of the expressions for numbers between *n* and the next higher base are formed by (a single) addition or subtraction of *n* or a multiple of *n* with expressions for numbers smaller than *n*.

This assumes that for any expression the linguist can unambiguously analyse each numeral expression into its constituent parts (or analyse it as consisting of only one part). As an example, for Swedish we would begin by finding the biggest part of the highest normed expression, which according to my own knowledge is *miljard* (10^9). Thereafter we can find the next lower base by trying divisors x of 10^9 to see if the numbers between x and 10^9 are expressed in the required form. E.g., $x = 5 \times 10^8$ is not because we do not say **en-halv-miljard plus ett* 'half-a-billion plus one' or the like for $5 \times 10^8 + 1$ or any, let alone a majority, of the numbers between 5×10^8 and 5×10^9. However, *miljon* (10^6) fulfils the requirements and we can easily arrive at the conclusion that Swedish has $\{10, 10^2, 10^3, 10^6, 10^9\}$ as its set of bases.

The definition of base as stated gives unambiguous decisions for formations which are sometimes (and sometimes not) called base by other authors; systematic subtractions, special lexemes for base-multiples, or isolated cases of addition, e.g., only 7 = 6 + 1 but otherwise no additions involving 6. Examples of such cases and their systematic resolution with my definition are given in Table 4.

Once the bases are known we can compactly describe each numeral expression in the multiplicative-additive form as:

$$a_n b_n + a_{n-1} b_{n-1} + \ldots + a_1 b_1 + a_0$$

Where b_i make up the set of bases, $a_i < b_i$ for all i, and $b_i > b_j$ if $i > j$. The a_i: s can be called coefficients and are uniquely determined for each numeral expression given the set of bases. This essentially matches all natural language numeral systems, since a non-small natural language numeral system which does not use bases is not known (the Appendix shows that a numeral system without base is logically possible). Moreover, the bases are always expressed as overt morphemes, in particular, they are never coded as place-values (an alternative way to say 11–19, . . ., 90–99 in spoken Samoan comes the closest (Mosel and Hovdhaugen 1992)). No further generalization over the formation of the b_i :s is possible because no system is known that uses exponentiation.[2] Occasionally it is a non-forced analysis to claim that $a_n < b_n$ for the highest base b_n. Pluses, and sometimes minuses, are expressed overtly or covertly and often different markings are used for different pluses between different bases.

---

2. Occasionally one may see e.g., 100 as 'big-ten' or so which may be called exponentiation, but still no case is known where this extends beyond one single form. Also, the Greek-derived words for 10^{15} and above in a lot of European languages are not (yet) common to the whole speech community.

Table 4. Examples of formation types and outcomes of the definition of base.

	Lutuami (Klamath-Modoc) (Dixon and Kroeber 1907: 673)		Nyokon (Niger-Congo) (Richardson 1957: 30)		Kare (Niger-Congo) (Dijkmans 1974: 147)		Ainu (Isolate) (Refsing 1986: 110)	
	Analysis	Expression	Analysis	Expression	Analysis	Expression	Analysis	Expression
1	1	nas	1	ámɔ̀	1	emotí	1	sine
2	2	lap	2	àfɔ́ɔ̀	2	ibili	2	tu
3	3	ndan	3	átár	3	etotu	3	re
4	4	umit	4	ɲnnɨs	4	biu	4	ine
5	5	tunip	5	ɲtɔ́ɔ̀r	5	etano	5	asikne
6	5 + 1	nas-ksapt	6	átʃɲ́n	5 + 1	etano na emoti	10 – 4	iwan
7	5 + 2	lap-ksapt	6 + 1	ɲtʃɲ́n námɔ̀	5 + 2	etano na ibili	10 – 3	arwan
8	5 + 3	ndan-ksapt	?	ɲyáá nɨ̀ màn	5 + 3	etano na etotu	10 – 2	tupesan
9	10 – 1	nas-xept	8 + 1	ɲyáá nɨ̀ màn námɔ̀	5 + 4	etano na bînu	10 – 1	sinepesan
10	10	te-unip	10	àwát	10	la-ato	10	wan
11	10 + 1	taunep-anta nas	10 + 1	àwát ámɔ̀	10 + 1	laäto na emoti	10 + 1	sine ikasma wan
…	…	…	…	…	…	…	…	…
15	…	…	…	…	15	sanga	…	…
16	…	…	…	…	15 + 1	sanga-na-emoti	…	…
…	…	…	…	…	…	…	…	…
20	2 × 10	lap-eni taunep	20	nɨ̀tʃɲ́n	2 × 10	atumbili	20	hot
21	2 × 10 + 1	lap-eni taunep-anta nas	20 + 1	nɨ̀tʃɲ́n ámɔ̀	…	…	20 + 1	sine ikasma hot
…	…	…	…	…	…	…	…	…
30	3 × 10	nda-ni taunep	3 × 10	àwát átár	2 × 10 + 10	atumbili na laato	20+10	wan e tu hot
…	…	…	…	…	…	…	…	…
40	…	…	…	…	…	…	2 × 20	tu hot
Base		5–10		10		5–20		5–10–20

One often sees that languages have irregularities between b_1 and $2 \times b_1$, for example the Spanish teens as shown in Table 5. Perhaps more often one sees irregularities in $x \times b_1$, especially $2 \times b_1$, particularly in the languages of Eurasia (e.g., Turkish (Lewis 2000) as in Table 5). Spanish and related varieties are virtually unique in having an idiosyncracy as high as in 500 – *quinientos* rather than the regular **cincocientos*.

Table 5. Irregularities in Spanish teens (left) and Turkish tens (right)

1	uno	11	once	1	bir	10	on
2	dos	12	doce	2	iki	20	yirmi
3	tres	13	trece	3	üç	30	otuz
4	cuatro	14	catorce	4	dört	40	kırk
5	cinco	15	quince	5	beş	50	elli
6	seis	16	dieciséis	6	altı	60	altmış
7	siete	17	diecisiete	7	yedi	70	yetmiş
8	ocho	18	dieciocho	8	sekiz	80	seksen
9	nueve	19	diecinueve	9	dokuz	90	doksan
10	diez	20	veinte	10	on	100	yüz

Generally Indo-European languages tend to have more irregularities in their numeral system than the impressionistic world average, perhaps culminating in Modern Indo-Aryan languages, e.g., Panjabi as above. Welsh (King 1993), though there are varieties which have restructured and switched to base 10, is another widely known case of a numeral system with many idiosyncracies.

It follows from the multiplicative-additive form that the most regular system is one where any plus is always expressed the same way and there are no other irregularities. Such a system is evidenced in e.g., Yamba (a Grassfields language of Cameroon) (Lucia 2001).

2.2 Complexity as global ordering constraints

If one assumes a description language using phrase-structure rules, which many people do, then not all rules are equally complex. Although completely regular, some phenomena require a more elaborate phrase structure grammar or even some more expressive description language. These complexities may be characterized as follows:

> The formation rule of a subpart of a numeral expression depends on the rest of the numeral expression.

Recall the multiplicative-additive expression decomposition from Section 2.1. We are now going to discuss the kind of complexity where a particular plus or a particular multiplication is expressed differently depending on the context in which it appears.

There are rare examples, e.g., Guajiro (Zubiri and Jusayú 1986), Kikongo (Söderberg and Widman 1966) and Breton (Press 1986), of languages with discontinuous numerals, i.e., when used attributively, they surround the noun that is quantified. This may also lead to cases where a numeral expression appears discontinuously within a larger numeral expression, such as the Breton 64 in (1).

(1) *pevar mil ha tri-ugent*
four thousand on three-twenty
'64 000'

Some languages, e.g., Erromangan (Crowley 1998), that have N Num order while still having big-before-small order between additive constituents get a potential ambiguity in phrases like 1000 8 which can then mean either 8000 or 1008.[3] Perhaps from pressure to resolve the ambiguity more clearly than by intonation alone, some languages have evolved reordering possibilities that depend on the diagnosed ambiguity of the whole resulting expression. Let's look at an example from Swahili which has N Num order generally. When a composite base multiplier and a rest are present, one can unambiguously use the order $b_i\, a_i + r$ (2).

(2) *elfu sabini na nane tatu*
thousand seventy and/with eight three
'78 003'

When the rest is not present the remaining expression may be ambiguous (3).

(3) *elfu sabini na nane*
thousand seventy and/with eight
'1 078' or '78 000'

One way to force the base-multiplier reading to put it before the base (4).

(4) *sabini na nane elfu*
seventy and/with eight thousand
'78 000'

This section shows that numeral systems can exhibit forms that are perfectly regular but require a longer description if a phrase-structure grammar is used as the description language.

2.3 Complexity vs. economy

Perhaps in response to a sudden demand for an expanded numeral system, there are languages which have evolved very morpheme-promiscuous number expressions e.g., Murinypata (Walsh 1976) in (5) or Maipure (Zamponi 2003) in (6).[4]

(5) *maŋenumimaŋenumimenumimenumimaŋenuminumi*
'hand-one-hand-one-foot-one-foot-one-hand-one-one'
'26'

3. N Num order is quite common (Dryer1989) whereas Malagasy (Parker 1883) seems to be the only modern language with consistent small-before-big ordering.

4. The abbreviations used in the gloss: CL classifier, NPOSS non-possessed.

(6) *papeta* *janà* *pauria* *capi-ti* *purenà*
one.CL.HUMAN take/follow(?) other hand-NPOSS relative
'one takes one relative from the other hand'
'6'

The forms have norms and are highly regular yet some grammar writers tend to label them "cumbersome" in comparison to the more familiar expressions in European languages. Presumably the alleged cumbersomeness of an expression lies in the duration, number of syllables or number of morphemes. In either case, we may speak of the *economy* of a numeral system as the average length (in duration, syllables or morphemes – I would opt for syllables) of its expressions. In particular, if two numeral systems have the same domain and one has longer expressions for every numeral in the domain, it is less economical. This label economical also preserves the intuition that cumbersomeness is something one wishes to spare. Most of the time grammars do not investigate the cumbersomeness in detail, but there are at least cases, e.g., in Murinypata, where "one native speaker with practically no formal education readily produced the number term for 'one hundred' which consists of seventy syllables" (Walsh 1976: 198).

Less economic numerals like these are uncommon, whereas numerals with about the same level of economy as English are standard. When we see cases where uneconomic numeral expressions are replaced with more economic numerals of another language, it is always the case that the language with the economic numerals is also socio-economically dominant. It is then sufficient to explain this numeral replacement within the general situation, i.e., take-over of the socio-economically dominant language, so a causal link between cumbersomeness (or the like) and abandonment is not necessary to explain these cases.

There appears to be no reason to link uneconomical numeral systems with complex ones. Indeed, the only difference between those and regular ones is the number of phonemes/syllables/morphemes. From a description length perspective, long forms are trivial to compress by code-booking as long as they are regular, so the addition in description length resulting for uneconomical numeral systems is negligible. Therefore, they will not be matter for further discussion.

2.4 Overall complexity

Probably the most complex numeral systems I have observed are those in Modern Indo-Aryan languages, like Panjabi. Breton does not have as many irregular items but probably has the most structural idiosyncracies; it is vigesimal up to about 200 then switches to decimal, it forms 50 and 150 with halves (half-hundred and hundred-half respectively), and has discontinuous formations.

Not counted are some numeral systems with enormous amounts of quite idiosyncratic forms when the difficulty invariably comes from the classifier fusing with the numeral, e.g., Ket (a Yeniseic language of Siberia) (Dul'zon 1968), rather than irregularity in forming the numerals themselves.

As for a "world average", there is neither time nor space to back it up with detailed evidence, but I may give an idea impressionistically. For these considerations, to make the objects comparable, imagine that, for all languages, we cut away all numerals over a hundred, and we leave out entirely languages that don't have numerals up to a hundred. Looking at all languages with a numeral systems up to a hundred this way, the average complexity of these would be similar to that of Camus above, i.e., around 15 forms to be learned by rote, one main formation rule and one formation sub-rule (no global ordering constraints). However, an average obtained from that language set will be less interesting because due to borrowing and inheritance the languages, and thus numeral systems, in question are not independent. If instead we look at the complexity of the subset of independent numeral systems – this time even more impressionistically – the average will come closer to he simplest possible, with around 11 rote forms and only one formation rule, on the same level as Nyokon above.[5]

3. Complexity in numeral systems of pidgin and creole languages

If the characterization of complexity makes sense, it would be highly interesting to see to what extent pidgins, pidgincreoles and creoles (definitions, following Bakker (2006), in Table 6) have the same amount or less complexity than their respective lexifier language. This question is immediately pertinent to the ongoing debate on the "simplicity" of creole languages (see e.g., McWhorter (2001) [also commentaries, pp. 167–412 in the same issue of the journal]). Since the alleged simplicity of creoles is held to be due to an earlier pidgin stage, the investigation of the numeral systems of all varieties of pidgins, pidgincreoles and creoles are relevant to the debate.

Table 6. Definitions of pidgins, pidgincreoles, and creoles

	Pidgins	**Pidgincreoles***	**Creoles**
has norms	yes	yes	yes
reduced from other language(s)	yes	yes	yes
ethnic or political group language	no	no	yes
native language	no	yes/no	yes
main language of speech community	no	no/yes	yes

*A pidgincreole should have a "yes" in at least one of the last two rows.

5. To be more precise, independent means that borrowing and inheritance can be ruled out. An independent set of numeral systems may be reached as follows. Exclude all systems which have morpheme(s) that have cognate(s) in some other family. Of the remaining systems, if two systems share at least one cognate put both of them in the same *chain*. Each chain is now independent from all other chains under the assumption that structural borrowing (metatypy) does not take place without simultaneous borrowing of a morpheme and that the etymology of numerals is generally known. To get an average one may take the average within each chain and then the average of the average of each chain.

In this study I have tried to gather data on numeral systems for all known pidgins, pidgincreoles and creoles, excepting only those whose status as such is doubted. However, a lot of descriptive accounts of pidgin/creole languages do not feature full data on numerals, and in a few more cases the publications containing the numeral data was not accessible to the author (and in an unknown number of cases a publication containing the sought information was simply not known to the author). The resulting set of pidgin/creole numeral systems are given in Table 7 together with a classification into pidgin (p), pidgincreole (pc) or creole (c) – acknowledging, however, that the sociohistorical data available to back up such a classification is quite uneven. To show whether restructuring of the numeral system of the lexifier language has taken place, I have grouped on the language that supplied the numerals. The language that supplied the numerals was usually the same as the lexifier for general vocabulary, but not always; e.g., in Fanakalo the numerals come from English but the general lexifiers are Nguni Bantu languages.

Most lexifiers had complexities in the numerals, allowing a test for the possibility of a simplification in the daughter pidgin/creole. A **C** in the teens/tens column indicates that there is some complexity in the formation, whereas an S indicates transparent formation, i.e., as $10 + x$ or $x \times 10$. No daughter language in our sample invented more complexity than the lexifier, except Kinubi which, unlike Arabic, has multiplier-base order for 10 and 100 but base-multiplier order for multiplications of 1000.

Table 7. Complexity in pidgin, pidgincreole and creole numeral systems

Language	Clf.*	Teens	Tens	Source
French		**C**	**C**	
Tayo	c	C	C	(Ehrhart 1993: 135)
Pointe Coupée	c	C	C	(Klingler 2003: 197–198)
Breaux Bridge	c	C	C	(Neumann 1985: 124)
Seychellois	c	C	C	(Bollée 1977: 39 + App. D)
Mauritian Creole	c	C	C	(Baker 1972: 144–145)
Haitian Creole	c	C	C	(Hall 1953: 29)
St. Lucian Creole	c	C	C	(Carrington 1984: 77–79)
Karipuna do Amapá	c	C	?	(Tobler 1987)
Tay Bôi	p	s	s	(Reinecke 1971: 52)
Arabic		**S**	**C**	
Nubi	c	S	C	(Luffin 2005: 157–167)
Turku	p	S	C	(Prokosch 1986: 100–101)
Portuguese		**C**	**C**	
Sãotomense	c	S	S	(Lorenzino 1998: 107–109)
Principense	c	C	C	(Günther 1973: 63–64)
Papia Kristang	c	C	C	(Baxter 1988: 48–49)
Crioulo Guiné	c	S	C	(Wilson 1962: 16–17)
Guinea-Bissau	c	S/C	C	(Honório do Couto 1994: 99–100)
Kap-Verde	c	C	C	(Veiga 1995: 172–173)
Annobonese	c	?	S	(Schuchardt 1888: 22)

(*Continued*)

Table 7. Continued.

Language	Clf.*	Teens	Tens	Source
Spanish		**C**	**C**	
Papiamentu	c	S	C	(van Name 1870: 154) (Munteanu 1996: 319–320)
Dutch		**C**	**C**	
Sranan	c	S	S	(Braun 2005: 295–307)
Negerhollands	c	C	C	(Oldendorp 1996 [1767–1768]: 58–154)
Ngbandi		**S**	**S**	(Kondangba 1991)
Sango	pc	S	S	(Giraud 1908: 266)
Russian		**C**	**C**	
Russenorsk	p	C	C	(Broch & Jahr 1981: 122)
Assamese		**S**	**S**	(Babakaev 1961)
Naga Pidgin	pc	S	S	(Bhattacharjya 2001: 143–146)
Chinook		**S**	**S**	(Ross 1849: 322–323)
Chinook Jargon	p	S	S	(Holton 2004: 69–70)
Choctaw		**S**	**S**	(Byington 1915)
Mobilian Jargon	p	S	S	(Drechsel 1997: 107)
MacKenzie Inuit		**S**	**S**	(Stefansson 1909: 232)
Eskimo Trade Jargon	p	S	S	(Stefansson 1909: 232)
English		**C**	**C**	
Broken (Torres Str.)	c	C	C	(Shnukal 1988: 252–253)
Saramaccan	c	C	C	(Wullschlägel 1965 [1854]: 14)
Coastal NG Pidgin	pc	S	S	(Laycock 1970: 1)
Highlands NG Pidgin	pc	S/C	S/C	(Wurm 1971: 81–82)
Tok Pisin	pc	S/C	S/C	(Verhaar 1995) (Mosel 1980: 61–63)
Bislama	pc	C	C	(Guy 1975: 26–28)
Solomons Pijin	pc	C	C	(Jourdan 2002)
Fanakalo	p	C	C	(Kaltenbrunner 1996: 92)
Samoan Plantation	p	S	S	(Mühlhäusler 1978: 100–101)
Krio	c	C	C	(Fyle & Jones 1980)
Carriacou	c	C	C	(Kephart 2000: 174–198)
Nigerian Pidgin	pc	C	C	(Faraclas 1996: 231)
Ndyuka	c	C	S	(Huttar & Huttar 1994: 532)
Cantonese Pidgin	p	C	?	(Hall 1944: 97)
Jamaican Creole	c	C	C	(Own Knowledge)
Cameroon Pidgin	pc	S	S	(Parkvall 2000: 107)
Hiri Motu	pc	C	C	(Wurm & Harris 1963)

*Clf. is short for classification; p = pidgin, c = creole, pc = pidgincreole.

A few remarks are in order. An entry like S/C means that the source gives parallel forms. In a few other cases there is a difference between an early and a late set of numerals, i.e., there has been borrowing (rarely internal restructuring) overlaying numerals attested earlier, in which case Table 7 shows only the earliest attested set. In another few more cases one may *suspect* borrowing, often from the lexifier language anew, but where I have found

no attestation of an earlier set, I analysed the attested forms – be they original or not. A couple of languages are curious in that they have numerals from two source languages. It was always possible to decide on a "main" lexifier for Table 7, but I could not discern whether the multi-source situations were original or the result after some borrowing.

3.1 Discussion

Whereas some pidgins/creoles do tend to analyticity, a majority do not. Parkvall (2000) shows similar findings for African creoles specifically. If it were necessary that pidgins, and/or languages descended from pidgins, have a maximally simple structure, we would have seen quite different empirical results. The prime example is Naga Pidgin which has a numeral system of the same complexity as Panjabi. However, it is also true that pidgins/creoles have slightly less complex numerals relative to their lexifiers. The same lexifier needs not produce the same result in its daughter pidgins/creoles. It appears that we can not predict where restructuring is more likely to take place if it takes place at all; for example Papiamentu restructures 500 to a regular formation, whereas the parallel Kap Verde and Guinea Bissau cases do not.

Impressionistically, the pidgin/creole numeral systems are on the average more complex than the world average, both if we count all systems or only independent systems. So it is not possible to look only at a numeral system and say whether it is from a pidgin/creole language or not. This appears to be easily explained by the fact that well-documented pidgins/creoles have a set of lexifiers which is non-representative of the (documented) languages of the world as a whole.

Furthermore, unless they borrow, languages which change from having only a few lexical numerals to a combinatorial system of numerals, universally do this by forming a 5-10-20 system with transparent formations. Pidgins do not follow this path even though they undoubtedly have the means sufficient to do so, i.e., juxtaposition and words for 'and', 'hand', 'foot' and 'man'.

Why is the prediction that pidgin/creole languages should be (maximally?) simple not borne out as regards numeral systems? The answer is obscure to me as lack of sociohistorical data prevent the mechanisms behind the prediction from being fully scrutinized.

References

Babakaev, V.D. 1961. *Assamskij Jazyk* [Jazyki Zarubezhnogo Vostoka i Afriki]. Moskva: Akademia Nauk SSSR.

Baker, P. 1972. *Kreol: A Description of Mauritian Creole*. London: C. Hurst.

Bakker, P. 2006. Pidgins versus other contact languages. In *Handbook of Pidgins and Creoles*, S. Kouwenberg & J.V. Singler (eds). Oxford: Blackwell.

Baxter, A.N. 1988. *A Grammar of Kristang (Malacca Creole Portuguese)* [Pacific Linguistics B 95]. Canberra: Australian National University.

Berger, H. 1992. Modern Indo-Aryan. In *Indo-European Numerals* [Trends in Linguistics. Studies and Monographs 57], J. Gvozdanovic (ed.), 243–288. Berlin: Mouton de Gruyter.

Bhattacharjya, D. 2001. *The Genesis and Development of Nagamese: Its Social History and Linguistic Structure*. PhD Dissertation, City University of New York.

Bollée, A. 1977. *Le créole français des Seychelles: Esquisse d'une grammaire - textes - vocabulaire* [Beihefte zur Zeitschrift für romanische Philologie 159]. Tübingen: Niemeyer.

Braun, M. 2005. *Word-Formation and Creolisation: The Case of Early Sranan*. PhD Dissertation, Universität Siegen.

Broch, I. & Jahr, E. 1981. *Russenorsk – et Pidginspråk i Norge* [Tromsø-Studier i Språkvitenskap III]. Oslo: Novus.

Byington, C. 1915. *A Dictionary of the Choctaw Language* [Bureau of American Ethnology Bulletin 46]. Washington: Smithsonian Institution.

Carrington, L.D. 1984. *St. Lucian Creole: A Descriptive Analysis of its Phonology and Morpho-Syntax* [Kreolische Bibliothek 6]. Hamburg: Helmut Buske Verlag.

Comrie, B. 1992. Balto-Slavonic. In *Indo-European Numerals* [Trends in Linguistics. Studies and Monographs 57], J. Gvozdanović (ed.), 717–833. Berlin: Mouton de Gruyter.

Crowley, T. 1998. *An Erromangan (Sye) Grammar* [Oceanic Linguistics Special Publication 27]. Honolulu: University of Hawaii Press.

de Vries, L.J. 1998. Body part tally counting and bible translation in Papua-New Guinea and Irian Jaya. *The Bible Translator (Practical Papers)* 49(4): 409–415.

Dijkmans, J.J.M. 1974. *Kare-taal: Lijst van woorden gangbaar bij het restvolk Kare opgenomen in de jaren 1927–1947*. Sankt Augustin: Anthropos-Institut – Haus Völker und Culturen.

Dixon, R.B. & Kroeber, A.L. 1907. Numeral systems of the languages of california. *American Anthropologist* 9(4): 663–689.

Drechsel, E.J. 1997. *Mobilian Jargon: Linguistic and Sociohistorical Aspects of a Native American Pidgin* [Oxford Studies in Language Contact]. Oxford: Clarendon.

Dryer, M.S. 1989. Large linguistic areas and language sampling. *Studies in Language* 13(2): 257–292.

Dul'zon, A.P. 1968. *Ketskij Jazyk*. Tomsk: Izdatel'stvo Tomskogo Universiteta.

Ehrhart, S. 1993. *Le Créole Français de St-Louis (Le Tayo) en Nouvelle Calédonie* [Kreolische Bibliothek 10]. Hamburg: Helmut Buske Verlag.

Faraclas, N. 1996. *Nigerian Pidgin* [Descriptive Grammars Series]. London New York: Routledge.

Fyle, C.N. & Jones, E.D. 1980. *A Krio-English Dictionary*. Oxford: Oxford University Press.

Giraud, G. 1908. Vocabulaire des dialectes sango, balkongo, et a-zandé. *Révue Coloniale, Nouvelle Serie* 58: 263–291, 332–354.

Günther, W. 1973. *Das portugiesische Kreolisch der Ilha do Príncipe* [Marburger Studien zur Afrika- und Asienkunde: Serie A 2]. Marburg an der Lahn.

Guy, J.B.M. 1975. *Handbook of Bichelamar/Manuel de Bichelamar* [Pacific Linguistics C 34]. Canberra: Australian National University.

Hafford, J.A. 1999. Elements of Wuvulu grammar. MA Thesis, University of Texas at Arlington.

Hall, Jr., R.A. 1944. Chinese Pidgin English grammar and texts. *Journal of the American Oriental Society* 64(3): 95–113.

Hall, Jr., R.A. 1953. *Haitian Creole: Grammar, Texts, Vocabulary* [Memoirs of the American Folklore Society 43]. The American Anthropological Association.

Hammarström, H. In preparation. Small numeral systems. Ms.

Heine, B. 1980. *The Non-Bantu Languages of Kenya* [Language and Dialect Atlas of Kenya II]. Berlin: Verlag von Dietrich Reimer.

Holton, J. 2004. *Chinook Jargon: The Hidden Language of the Pacific Northwest*. San Leandro, California: Wawa Press.

Honório do Couto, H. 1994. *O Crioulo Português da Guiné-Bissau* [Kreolische Bibliothek 14]. Hamburg: Helmut Buske Verlag.

Huttar, G.L. & Huttar, M.L. 1994. *Ndyuka* [Descriptive Grammars Series]. London New York: Routledge.

Jourdan, C. 2002. *Pijin: A Trilingual Cultural Dictionary* [Pacific Linguistics 526]. Canberra: Australian National University.

Kaltenbrunner, S. 1996. *Fanakalo: Dokumentation einer Pidginsprache* [Beiträge zur Afrikanistik 53 / Veröffentlichungen der Institute für Afrikanistik und ägyptologie der Universität Wien 72]. Vienna: Afro-Pub.

Kephart, R.F. 2000. *"Broken English": The Creole Language of Carriacou* [Studies in Ethnolinguistics 6]. New York: Peter Lang.

King, G. 1993. *Modern Welsh: A Comprehensive Grammar*. London: Routledge.

Klingler, T.A. 2003. *If I Could Turn My Tongue Like That: The Creole Language of Pointe Coupée Parish, Louisiana*. Baton Rouge: Louisiana State University Press.

Kondangba, Y. 1991. Structure des numeraux en bantu (lingɔmbɛ) et en non-bantu (ngbaka, minagende, ngbandi, ngbundu, mɔnɔ, mbanza). *Annales Æquatoria* 12: 307–319.

Laycock, D.C. 1970. *Materials in New Guinea Pidgin (Coastal and Lowlands)* [5*Pacific Linguistics: Series D*]. Canberra: Australian National University.

Laycock, D.C. 1975. Observations on number systems and semantics. In *New Guinea area languages and language study, vol 1: Papuan Languages and the New Guinea linguistic scene* [Pacific Linguistics C 38], S.A. Wurm (ed.), 219–233. Canberra: Australian National University.

Lean, G.A. 1992. *Counting Systems of Papua New Guinea and Oceania*. PhD Dissertation, Papua New Guinea University of Technology.

Lewis, G.L. 2000. *Turkish Grammar* (2nd ed.). Oxford: Oxford University Press.

Lorenzino, G.A. 1998. *The Angolar Creole Portuguese of São Tomé: Its Grammar and Sociolinguistic History* [LINCOM Studies in Pidgin & Creole Languages 1]. München: Lincom Gmbh.

Lucia, N.N. 2001. Yamba: A morphosyntactic study of the basic sentence. MA Thesis, The University of Yaoundé I.

Luffin, X. 2005. *Un créole arabe: Le kinubi de Mombasa, Kenya* [LINCOM Studies in Pidgin & Creole Linguistics 07]. München: Lincom GmbH.

McGregor, W.B. 2004. *The Languages of the Kimberley, Western Australia*. London New York: Routledge.

McWhorter, J. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology* 5(3/4): 125–166.

Mosel, U. 1980. *Tolai and Tok Pisin: The Influence of the Substratum on the Development of New Guinea Pidgin* [Pacific Linguistics B 73]. Canberra: Australian National University.

Mosel, U. & Hovdhaugen, E. 1992. *Samoan Reference Grammar* [Instituttet for sammenlignende kulturforskning 85]. Oslo: Scandinavian University Press.

Mühlhäusler, P. 1978. Samoan Plantation Pidgin English and the origin of New Guinea Pidgin. In *Papers in Pidgin and Creole Linguistics 1* [Pacific Linguistics A 54], 67–120. Canberra: Australian National University.

Munteanu, D. 1996. *El Papiamento, lengua criolla Hispánica* [Biblioteca Románica Hispánica: Tratados y Monografías 17]. Madrid: Editorial Gredos.

Neumann, I. 1985. *Le Créole de Breaux Bridge, Louisiane* [Kreolische Bibliothek 7]. Hamburg: Helmut Buske Verlag.
Oldendorp, C.G.A. 1996 [1767–1768]. *Criolisches Wörterbuch* [Lexicographica: Series Maior]. Tübingen: Max Niemeyer.
Parker, G.W. 1883. *Concise Grammar of the Malagasy Language*. London: Trübner & Co.
Parkvall, M. 2000. *Out of Africa: African Influences in Atlantic Creoles*. London: Battlebridge Publications.
Press, I. 1986. *A Grammar of Modern Breton* [Mouton Grammar Library 2]. Berlin: Mouton de Gruyter.
Prokosch, E. 1986. *Arabische Kontaktsprachen (Pidgin- und Kreolsprachen) in Afrika* [Grazer Linguistische Monographien 2]. Graz: Institut für Sprachwissenschaft der Universität Graz.
Refsing, K. 1986. *The Ainu Language: The Morphology and Syntax of the Shizunai Dialect*. Aarhus: Aarhus University Press.
Reinecke, J.E. 1971. Tay bôi: Notes on the Pidgin French of Vietnam. In *Pidginization and Creolization of Languages: Proceedings of a Conference Held at the University of the West Indies, Mona, Jamaica, April 1968*, D. Hymes (ed.), 47–56. Cambridge: Cambridge University Press.
Richardson, I. 1957. *Linguistic Survey of the Northern Bantu Borderland* [Linguistic Survey of the Northern Bantu Borderland 2]. Oxford: Oxford University Press.
Ross, A. 1849. *Adventures of the First Settlers on the Oregon or Columbia River*. London: Smith, Elder & Co.
Schuchardt, H. 1888. *Ueber das Negerportugiesische von Annobom* [Kreolische Studien VII]. Wien: Carl Gerold's Sohn.
Shackle, C. 2003. Panjabi. In *The Indo-Aryan Languages* [Routledge Language Family Series], G. Cardona and D. Jain (eds), 581–621. London/New York: Routledge.
Shnukal, A. 1988. *Broken: An Introduction to the Creole Language of Torres Strait* [Pacific Linguistics C 107]. Canberra: Australian National University.
Söderberg, B. & Widman, R. 1966. *Kikongo*. Stockholm: Svenska Bokförlaget Bonniers.
Stefansson, V. 1909. The Eskimo trade jargon of Herschel Island. *American Anthropologist* 11(2): 217–232.
Tobler, A.W. 1987. *Dicionário crioulo karipúna/português, português/crioulo karípúna*. Brasília: Summer Institute of Linguistics.
van Name, A. 1869–1870. Contribution to creole grammar. *Transactions of the American Philological Association* 1: 123–167.
Veiga, M. 1995. *O Crioulo de Cabo Verde: Introdução à Gramática* (2nd ed.). Praia: Instituto Caboverdiano do Livro e do Disco, Instituto Nacional da Cultura.
Verhaar, J.W.M. (ed.) 1995. *Toward a Reference Grammar of Tok Pisin: An Experiment in Corpus Linguistics* [Oceanic Linguistics Special Publication 26]. Honolulu: Univeristy of Hawai'i Press.
Vitányi, P. & Li, M. 1997. *Kolmogorov Complexity.* (2nd ed.). Berlin: Springer-Verlag.
Walsh, M.J. 1976. *The Murinypata Language of North–West Australia*. PhD Dissertation, Australian National University.
Wilson, W.A.A. 1962. *The Crioulo of Guiné*. Johannesburg: Witwatersrand University Press.
Wullschlägel, H.R. 1965 [1854]. *Kurzgefasste Neger-Englische Grammatik*. Amsterdam: S. Emmering.
Wurm, S.A. 1971. *New Guinea Highlands Pidgin: Course Materials* [Pacific Linguistics D 3]. Canberra: Australian National University.

Wurm, S.A. & Harris, J.B. 1963. *Police Motu: An Introduction to the Trade Language of Papua (New Guinea) for Anthropologists and Other Fieldworkers* [Pacific Linguistics B 1]. Canberra: Linguistic Circle of Canberra Publications.

Zamponi, R. 2003. *Maipure* [Languages of the World/Materials 192]. München: Lincom GmbH.

Zubiri, J.O. & Jusayú, M.A. 1986. *Gramática de la Lengua Guajira (Morphosintaxis)*. San Cristóbal: Universidad Católica del Tachira.

Appendix: Logically possible numeral systems without base

By a numeral system I mean a finite set of atoms used combinatorially to denote each member of a serially ordered target set. If the set of atoms has cardinality n, each combinatorial expression may not be longer than $2n$, and the target set must have cardinality of at least 2^n.

There are several ways in which one can have a numeral system without a base (as defined in section 4). I will sketch a few examples here.

Example 1. Three Alterating Bases: Let the set of atoms be A = {1,2,3,4,5,6,7,8,9, 10,11,12,100,110,120,10000,11000,12000}. Given a number n to be expressed, decompose $n = pq + r$ such that $p < q$ and $q \epsilon A$ is maximal. Then express n as "p × q + r" with possible recursion until $p \epsilon A$. So e.g., 31 as 3 × 10 + 1, 34 as 3 × 11 + 1, 121 as 120 + 1, 778 as 7 × 110+8,132012001 as 11000 + 1 × 12000 + 1 and so on.

Example 2. Decomposition into Primes: The fundamental theorem of arithmetic says that every number n can be written as a product of primes $p_1^{e_1} \cdots p_n^{e_n}$. Thus we can have the primes as our set of atoms and express any $n > 1$ as: $\mathrm{E(n)} = p_1(E(e_1))p_2(E(e_2))\ldots p_n(E(e_n))$. Of course, with E(1) = 1.

Example 3. Increasing Gaps: Instead of letting counting begin anew at uniform intervals we can have some more complex evolution of intervals. For example, instead of re-counting at 10, 20, . . ., 100 etc we can increase the gap size at each step, e.g., 10, 21, 33, and so on.

Example 4. Permutations: For a completely non-transparent counting system we might use permutations. With the set of atoms 1–9 we can form the set of permutations of the numbers 1–9. This set can be ordered using the number obtained by reading the permutation as a place-value number expression. So e.g., 123456789 is the smallest, and would be used to represent 1, 123456798 is 2 and so on.

Example 5. Subsets: Also using the atoms 1–9 we can denote numbers by subsets of 1–9. An ordering of the subsets that does not yield the existence of bases is the following. Each pair of subsets of the same cardinality can be compared in terms of what I shall call "smallness" – the sum of its members, and if that is a tie the set with the smallest member that is not present in the other. Now, to map subsets to numbers, first take the smallest one-member subset, then the smallest two-member subset, . . ., nine-member subset, the next smallest one-member subset and iterating so on until all the sizes of subsets are exhausted. This will yield, in increasing order, {1} {1,2} {1,2,3} {1,2,3,4} . . . {1,2,. . . ,9} {2} {1,3} {1,2,4} {1,2,3,5} {1,2,3,4,6} . . . .

Explaining Kabuverdianu nominal plural formation

Angela Bartens & Niclas Sandström
University of Helsinki

In this study, we apply the morphosyntactic 4-M model developed by Myers-Scotton and Jake (2000a, b) to data from Kabuverdianu or Cape Verdean Creole Portuguese (CVC) which has been less strongly restructured than so-called prototypical creoles.[1] We focus on nominal plural marking where CVC presents similar morphosyntactic configurations as Brazilian Vernacular Portuguese (BVP) and the Portuguese spoken by the "Tongas". Previous accounts of CVC and BVP nominal plural marking mention the occurrence of (at least) one inflectional marker per NP. We argue that the reduction of inflectional plural marking in CVC constitutes a case of overall loss of morphosyntactic complexity which is due to CVC having arisen through substantial reduction and restructuring during creolisation and to having shallow time-depth of existence in comparison to older languages, e.g., its lexifier Portuguese. We also argue that 4-M theory may constitute a useful diagnostic tool for the prediction of the configurations of complexity vs. simplification in cases of language reduction.

1. Introduction

The aim of this paper is to discuss Kabuverdianu nominal plural marking strategies from the point of view of linguistic complexity. At the same time, this is one of the first attempts at applying the 4-M theory, which we will briefly introduce below, to a Creole or restructured variety. In consequence, it is also an attempt at testing the usefulness of the 4-M theory as a diagnostic tool for predicting the configurations of relative complexity in cases of language reduction and restructuring such as Creole formation to which we consider pidginisation to be an obligatory prerequisite (cf. McWhorter 2005: 11). We will present different plural marking strategies in Cape Verdean Creole Portuguese (henceforth, CVC), mainly those representative of the variety spoken on Santiago. Since Brazilian Vernacular Portuguese (BVP) and the restructured Portuguese spoken by the

1. We would like to thank the editors of this volume and two anonymous reviewers for their helpful comments. The usual disclaimers apply.

"Tongas" (see below) employ nominal pluralisation strategies which at least at first sight closely resemble those of CVC, nominal plural marking in these two varieties will also be considered for the sake of comparison. This entails a brief appreciation of previous studies of the non-realization of plural agreement in Portuguese varieties.

We find it useful to add that plural marking was chosen as the feature to be studied as diagnostic of linguistic complexity not only on the basis of the mentioned parallels with plural marking phenomena in other Portuguese varieties but also because inflectional morphology such as nominal pluralisation morphology has been found to provide a good testing case for the measuring of cross-linguistic complexity (cf. Kusters 2003: 18).

2. Kabuverdianu or Cape Verdean Creole Portuguese

Kabuverdianu is a Portuguese-based Creole language spoken by over a million Cape Verdeans. Less than half of these Kabuverdianu speakers live on the Cape Verde Islands (483,000 inhabitants in 2003). The remainder lives in a diaspora divided between the United States and several other countries. The Cape Verdean language and culture form ties which hold the diaspore together.[2]

The Cape Verdean Islands consist of nine inhabited islands. There is considerable linguistic variation between the different varieties and it can be assumed that a partly autonomous creolisation process took place on each inhabited island (Bartens 2000). These differences do not, however, constitute any obstacle to intercomprehension. In this study, we focus mainly on the variety spoken on Santiago, historically the first and, thence, most "original" variety in the Cape Verdean cluster to have emerged and one which has considerable demographic and political weight in the archipelago.[3]

The Cape Verde islands were discovered between 1456 and 1460. Settlement of the main island Santiago was initiated during 1460. São Vicente, today the most densely populated island in the north, was settled only in 1794. As a rule, the northern islands were settled later and with fewer slaves. In spite of different settlement patterns, the social interaction between colonizers and slaves was quite intense on all Cape Verde islands. Among the relatively few immigrants from Portugal, there was a substantial majority of males which in turn led to a considerable mixing of races. At present, 74% of the population of the archipelago is considered to be of mixed origin, the rest being divided between 3% whites and 23% blacks.

---

2. Cf. Veiga (2005) on the "global Cape Verdean nation".

3. Santiago has approximately 236,000 inhabitants and the nation's capital Praia is located there. Especially the town of Mindelo on São Vicente has disputed Praia's role as the main city of the archipelago. This has lead to linguistic antagonism as well (Bartens 2000:43).

Due to settlement history, the southern Sotavento varieties not only crystallized earlier but they also show more substrate influence and are more basilectal on a hypothetic Creole continuum than the northern Barlavento varieties.[4] This notwithstanding, even the variety of Santiago is less strongly restructured than so-called prototypical Creoles. The other varieties, including those of Barlavento, should be considered dilutions of that of Santiago due to the fact that the Santiago variety was taken to the other islands as part of the superstrate, leading to partially new creolisation processes. All these later variants would thus be second-generation Creole variants, the older ones being first-generation variants (with respect to Santiago).[5] Differences occur not only from island to island, but also within islands, although Cape Verdean dialectology is still in its infancy. It is also possible that the presumed dialectal variation within a single island is rather diastratic than diatopic in the sense that apparent geographic lects have arisen through social stratification of neighbourhoods, not through geographic distance by itself (Meintel 1975: 237–238).

The present-day situation has been characterized both as functional bilingualism (Thiele 1991: 28, 35) and as diglossia where bilingualism merely constitutes a future goal (Veiga 1995: 29–35). Creole language promotion seems to be gaining impetus only now, although Kabuverdianu was declared the national (albeit not official) language of the newly independent country over 30 years ago.

3. The 4-M model and creolization

The foundations of the 4-M theory lie in the Matrix Language Framework presented in Myers-Scotton (1993). The 4-M model itself is presented in Myers-Scotton and Jake (2000a, b). The authors claim that 4-M theory is a universal model for the classification of morphemes according to their degree of cognitive activation in the process of utterance formation. As the name of the model suggests, four different categories of morphemes are distinguished, content morphemes and three types of system morphemes.

According to the 4-M model, language production is postulated to pass through four different levels of activation. Speakers' intentions initially start forming at the conceptual level. These intentions activate language-specific semantic/pragmatic feature bundles which in turn select lemmas or language-specific entries on the level of the mental

4. Cf., e.g., Rickford (1988) for the notion of a Creole continuum which had been previously discussed by Labov, Bailey, DeCamp, and Bickerton in the 1970s.

5. The notion of first- and second-generation Creoles was first introduced by Chaudenson (1992). The Creole of São Nicolau has been traditionally counted as a Barlavento variety. Cardoso (1989: 17–18) argues that the Creole of São Nicolau occupies an intermediate position between the Sotavento and Barlavento varieties.

lexicon, also called lemma level. The lemmas mediate between the intentions of the conceptual level and the production of grammatical structures, including surface structures, at the two later levels, the functional level and the positional level, by sending directions to an abstract unit postulated in 4-M theory, the formulator, which turns on the actual morphosyntactic and morphophonological procedures that eventually result in surface-level utterances. Two kinds of lemmas are directly elected at the conceptual level: those which underlie content morphemes and those which underlie early system morphemes. Content morphemes can be identified by the fact that they are able to assign or receive thematic roles. Early system morphemes do not express contents *per se* and therefore do not assign or receive thematic roles but they are nevertheless needed already at the conceptual level to express speakers' intentions. Two more types of system morphemes become salient at the functional level where larger constituents are assembled, therefore called late system morphemes: so-called bridge and outsider late morphemes. The difference between the two types of late system morphemes depends on whether they refer to grammatical information within or outside of the immediate maximal projection in which they occur. Finally, the utterance is given phonetic shape at the positional level (Myers-Scotton & Jake 2000a). Examples of the different types of system morphemes from English are the following: the definite article *the* is an early system morpheme, the preposition *of* is a bridge morpheme, and third person singular *–s* in verbs is an outsider system morpheme (Myers-Scotton and Jake 2000a: 1063–1064).

Myers-Scotton (and associates) have attempted to extend the scope of the model to the creolisation context as well. As far as the application to Creole genesis is concerned, Myers-Scotton (2001: 239, 242, 251–253) argues that the substrate languages provide an abstract, compositional Matrix Language. The structure of the emerging Creole is supposedly mapped onto this abstract frame. The lexifier or superstrate language provides the bulk of the content morphemes.[6] System morphemes are generally lost, with the exception of those superstrate early system morphemes which are satellites to heads. In addition, superstratal content morphemes may be reanalyzed as system morphemes in the process of creolisation.

4. Nominal plural marking in Portuguese and restructured varieties of Portuguese

Definiteness, an early system morpheme according to the 4-M model, is in Portuguese linked to the expression of the notion of plurality. This joint marking of plurality and definiteness occurs in an obligatory manner on the first element of the NP to the left

6. In most Creoles, content morphemes retained from the substrate languages amount to only a few percent of the total.

of the nucleus (cf. Lopes 2005: 76). In addition, Standard European Portuguese (SEP) shows number agreement on every item either preceding the NP (that is, to the left of the nucleus) or following it (to the right of the nucleus).

(1) *Isto importa para a reconstituição*
this matters to DEF-FEM-SG reconstruction-SG
do-s ***dialecto-s*** *antigo-s*.
of.DEF-MASC-PL dialect-PL old-PL
'This is important to the reconstruction of the old dialects.'
(Castro 2004: 94; adapted)

(2) *a-s grande-s* ***compilaçõe-s*** *editoria-is iniciada-s*
DEF-FEM-PL big-PL compilation-PL editorial-PL initiated-FEM-PL
no século XIX
in.DEF-MASC century-SG 19th
'the extensive editorial compilations initiated in the 19th century'
(Castro 2004: 98; adapted)

In (1), we see the form *dos*, an amalgam of the preposition *de* and the masculine plural definite article *os* (*de* + *os* > *dos*). What is important in both the examples is that each element, in addition to the nucleus, either preceding or following the nucleus (written in boldface), is in agreement with the nuclear NP. The number of pre- or postposed elements is basically only limited by how many sequential lexemes can be perceived as intelligible.

Standard Brazilian Portuguese (SBP) is similarly sensitive to marking all determiners, nouns and adjectives of a noun phrase for plural number (Holm 2003: 101). Brazilian Vernacular Portuguese (BVP), in turn, tends to reduce plural marking, adding the plural suffix {-s} to only one element in the noun phrase. This element is most often a determiner, and the possible following nouns and adjectives are optionally inflected for the plural. In the literature, the suffix is often said to be realized only on the first element of the NP, frequently attributed to the fact that plurality is marked at the beginning of the NP in many Niger-Congo languages (Holm 2003: 101–102). The element marked need, however, not be the initial one, a configuration first noticed by Scherre (1988). This is the case for instance in examples (3) and (4), the latter one from (Holm 2003: 102).

(3) *na minha-s* ***coisa***
in.DEF-FEM-Ø my-FEM-PL thing-FEM-Ø
'in my things/in the things of mine'

(4) *todo os mais* ***velho***
all-Ø DEF-MASC-PL more old-MASC-Ø
'all the oldest (ones)'

The plural marking of BVP as such resembles to a surprising degree that of CVC:

(5) *varius* ***koléga***
various-PL colleague-Ø
'various colleagues'

(6) *tudu* *kes* ***trabajador*** stranjeru
all that-PL worker-Ø foreigner-Ø
'all those foreign workers'

(7) *kél* *dôs* ***pronomi***
that-SG two pronoun-Ø
'those two pronouns'

In (5)–(7), plurality is explicitly marked only on the determiner which is immediately adjacent to the nucleus (in italics) on the left.

Lopes (2005) demonstrates that the patterning of BVP nominal plural marking can be explained with the help of 4-M theory. Plural marking through suffixation of {-s} is favoured very strongly on the nucleus of the NP if the nucleus is not preceded by any determiner. If there are preceding determiners, marking on the element immediately adjacent to it on the left is favoured. She explains this as being a result of the fact that in these instances, the suffix {-s} is an early system morpheme while all the other suffixes which potentially occur in Standard Brazilian and European Portuguese are late system morphemes. She argues that in the acquisition of Portuguese through speakers of other languages as has massively occurred during the colonial history of Brazil, early system morphemes are perceived as relevant and retained while late system morphemes are not acquired.

This convincing account is paralleled by the findings of Baxter (2004) who discusses nominal plural formation patterns in the Portuguese of the so-called "Tongas", descendants of ex-slaves of African ancestry living on São Tomé whose L1 Portuguese bears the imprint of the L2 Portuguese input of previous generations and of language contact with and transfer from Umbundu, a southern Angolan Bantu language, now spoken in a koineised form among the Tongas studied by Baxter. Note that this language contact situation is very similar to the one which prevailed in colonial Brazil where the main African contact languages were the Angolan Bantu languages Kimbundu and, to a lesser degree, Umbundu.[7] As far as the linear order of the constituents of an NP is concerned, Baxter notes that the position at the immediate adjacent left of the nucleus prominently favours overt plural marking. In the oldest age group (61+ years), a generation with L2 Portuguese speaking parents, this prominence reaches a statistical significance between 0.95 and 0.97 (2004: 118). Baxter relates the prominence of this strategy which slightly diminishes for the younger age groups (the statistical significance of the immediate left position oscillates between 0.89 and 0.90 in the 41–60 years age group and is constant at 0.83 in the 20–40 group; Baxter 2004: 118) to the importance of plural marking through prefixation in Umbundu and other languages

7. This is the big picture: regionally and over shorter periods of time languages from Western Africa predominated, e.g., Fongbe well into the 18th century in Minas Gerais and Yoruba in 19th century Bahia.

and suggests (2004: 121) that in terms of Myers-Scotton's framework, Portuguese L2 content morphemes would have been inserted into an Umbundu L1 frame in which plural marking is required at the immediate left of the nucleus.

Considering the similarities between plural marking strategies in BVP and the Portuguese spoken by the Tongas, two restructured varieties of Portuguese, on the one hand, and CVC on the other, we posed ourselves the question whether the 4-M model could be applied to Kabuverdianu as well.

5. Previous descriptions of nominal plural marking in CVC

Descriptions of CVC plural marking insist on the restricted use of the superstratal suffix {-s}.[8] Lang (2002: xxx) states that it is suffixed to the first word "que admite a sua agregação" ('which allows for it to be added'; note the broad definition) and may be omitted altogether if the plurality can be gleaned from the context, a feature typical of Creole language pluralisation strategies in general. Quint (2000: 323–324) lists a number of restrictions (numerals, quantifiers, semantics, general context) and states that the suffix attaches only to the first element in the NP, an analysis frequently given for BVP as well as we just saw. Veiga (2000: 141) considers suffixation to be rare in this context and to occur only whenever it is not possible to use quantifiers. Baptista (2002: 35–42) emphasizes the importance of human/animate and definiteness as factors which strongly favour the suffixation strategy.

Brito (1967: 348), writing in the 1880s, gives only examples involving numerals when discussing CVC nominal plural formation.

6. CVC plural marking in our data

We gathered CVC data from various sources (see References) and compiled a pilot corpus of 500 NPs.[9] NPs containing possessive pronouns were eliminated since they may show number agreement with the NP they modify (cf. Baptista 2002: 59). Cf.

---

8. Singulars ending in a consonant take the allomorph {-is}, e.g., *mudjer* 'woman', *mudjeris* 'women' (cf. Lang 2002: xxx). This phonetically conditioned allomorphy is not discussed in much detail in the literature on CVC and it might be worthwhile to investigate whether this division of labour is realized in such a categorical manner as Lang (p.c.) claims, cf. e.g., *armun* 'brother/sister', *armunsis* 'brothers/sisters'. However, our main concern in this paper is to contrast presence and absence of morphological nominal plural marking.

9. We agree with one of the anonymous reviewers that a larger corpus should be compiled in the future. Other possibilities for expanding this study would be looking at CVC plural allomorphy (cf. note 8) and submitting the data to VARBRUL analysis as was done by Lopes (2005) and Baxter (2004).

(8) *si* ***kasa***
POSS-SG house-SG
'his/her house' (one owner, one house)

(9) *ses* ***kasa***
POSS-PL house-SG
'their house' (several owners, one house)

(10) *si* ***kasas***
POSS-SG house-PL
'his/her houses' (one owner, several houses)

(11) *ses* ***kasas***
POSS-PL house-PL
'their houses' (several owners, several houses)

There are some frequently recurring prenominal determiners in CVC: *tudu* 'all, every, everything' (CVC) which has both singular/collective and plural readings as it is an amalgam of the meanings of *todo/-a* 'all, every, the whole of, any' (Port.) and *tudo* 'all, everything' (Port.); *uns* (CVC) and *alguns* (CVC), both 'some' (plural) < *uns, alguns* (Port.); *kel* 'this' (CVC) and *kes* (CVC) 'these' < *aquel, aqueles* (Port.); *otu* 'other' (CVC) and *otus* 'others' (CVC) < *outro(s)* (Port.), and, of course, numerals.

The essential test for identifying content morphemes as opposed to any kind of system morphemes in 4-M theory is whether they can assign or receive thematic roles. This is true of all the Portuguese items which have been recruited into CVC as the determiners just mentioned. Applying the 4-M model to our data, we consider that these superstratal content morphemes are recruited as – or may assume the function of – early system morphemes such as non-redundant plural markers in CVC. The fact that *kel – kes* and *otu – otus* vary according to number does not invalidate this claim as it can be argued that the variation is lexicalized.[10]

As stated above, superstratal system morphemes can be retained according to the 4-M theory only if they are satellites which are accessed along with their heads. This would allow for those Portuguese plural suffixes to have been retained in CVC which are suffixed to the nucleus of the NP, provided plurality is not already expressed by a preposed determiner.

Let us now look at our data. Table 1 shows the predominant patterns found in our CVC data:[11]

---

10. In the late 19th century, *otu* could be pluralized only with the help of a numeral (Brito 1967: 358).

11. Very infrequently, a few other patterns occurred, e.g., *tantus ótu múzikus* 'so many other musicians' (VNG), where plurality was marked on the non-adjacent determiner and on the nucleus but not on the adjacent determiner; postposed *tudu* 'all; everything' did not seem to affect plural marking on the preceding NP in a systematic way (see ex. 27).

Table 1: Predominant patterns of CVC nominal plural marking.

–2	–1	nucleus	+1	values/ position	type	Conformity 4-M model
	tres	fiju		+ –	I	yes
	uns	saltiador		+ –	I	
kel	otus	rapariga		– + –	I	
	kes	mininu	fémia	+ – –	I	
		instrumentus	pedagójiku	+ –	II	
kes	tres	irmás		+ + +	III	no
tudu	kes	kumidas		+ + +	IV	
		kusa	nobu	– –	V	
kes	ótu	omi		+ – –	VI	
		osu	tudu	– +	VII	

Elements to the immediate adjacent left of the nucleus are given in column –1 and elements preceding the immediate adjacent left position in column –2. Elements following the nucleus are given in column +1. As examples with more than one element following the nucleus of the NP appear to be virtually inexistent in CVC, no column +2 was included in the table. In the values/positions column, [+] stands for plural marking, [–] for the absence of it. The value [+/–] which stands for the nucleus of the NP is given in bold print in order to make it easier to map the values sequence onto the examples given in the columns to the left. Examples of the structures, coded as types I–VII, will be introduced in the text below. Note that types I and II conform to the predictions of the 4-M model and that types III–VII contradict them. Figure 1 below further presents the patterns.

As could be verified from the data, the overt plural marking is most frequently expressed on the *first* element to the immediate left of the nucleus whenever this slot is filled by an element. The arrows in Figure 1 show the abstract mechanism, transforming the –1 element into the carrier of overt plural information. The dotted arrow stands for the "intermediate" level on which the overt plurality eventually moves to the left, leaving the nucleus unmarked. The brackets merely present the slots that most frequently would be filled with content morphemes in the CVC NP.

The majority of NPs patterns as predicted by the 4-M model. One option is marking plurality on the determiner immediately adjacent to the left of the nucleus of the NP (259 of the instances, i.e., 51.8 %). This is coded as type I in the table, e.g.:

(12) *kel* *otus* ***rapariga*** (B 63)
that-SG other-PL girl-Ø
'the other girls'

(13) *kel* *otus* ***pais*** (NK 84)
that other-PL country-Ø
'those other countries'

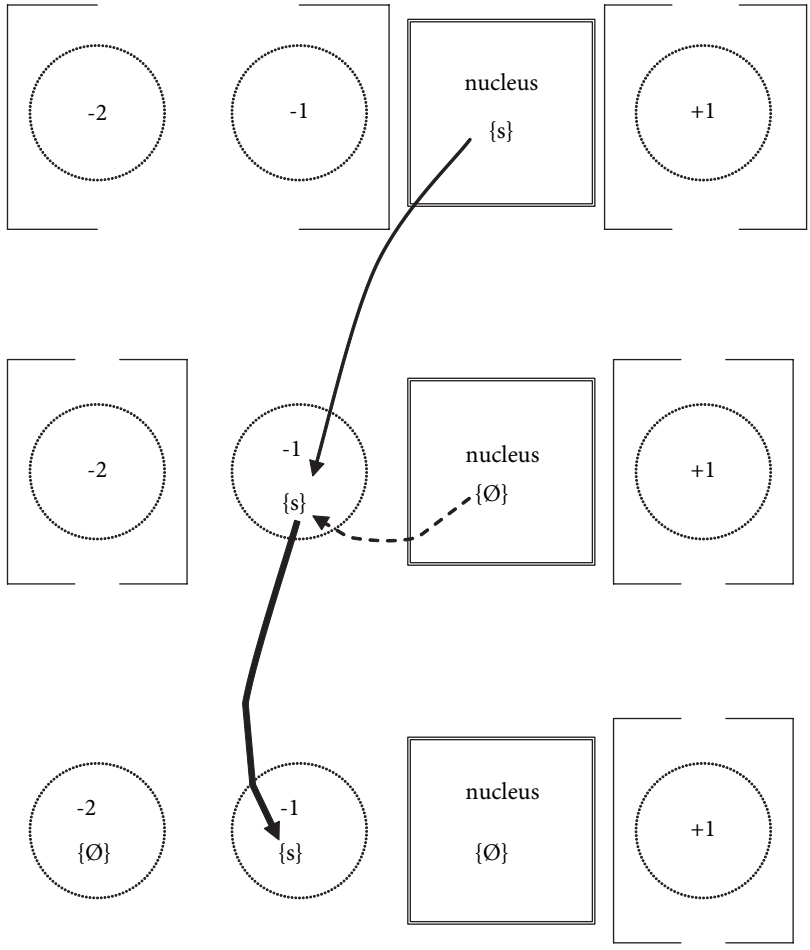


Figure 1: Predominant patterns of CVC nominal plural formation.

(14) *otus* ***lingua*** *importanti* (VKK)
other-PL language-Ø important-Ø
'other important languages'

Alternatively, {-s} may be suffixed to the nucleus (type II) as there are no preceding determiners (172 cases, 34.4 %), e.g.:

(15) ***obujetibus*** *fundamental* (VKK)
object-PL fundamental-Ø
'fundamental objects'

(16) ***familias*** *pobri* (NK 95)
family-PL poor-Ø
'poor families'

(17) ***paredis*** *bunitu* (B 25)
wall-PL beautiful-Ø
'beautiful walls'

(18) ***kulturas*** (VNG)
culture-PL
'cultures'

As Baptista (p.c.) states, the feature human/animate strongly favors the suffixation of plural {-s} to CVC NPs. We believe that this is due to some kind of resistance to a process of transformation of Portuguese singular object nouns into CVC set nouns (see below). The human/animate factor seems to account for the 28 instances (5.6 %) where double-marking of plurality (type III) occurred, e.g.:

(19) *kes* *tres* ***rapasis*** (B 80)
that-PL three boy-PL
'those three boys'

(20) *kes* *tres* ***irmás*** (B 101)
that-PL three sister-PL
'those three sisters'

Definiteness, the other factor found by Baptista (p.c.) to favor {-s} suffixation and which we likewise see in terms of resistance to a large-scale transformation of singular object into set nouns, could account for seven more instances (i.e., 1.4 %) of double-marking (type IV), e.g.:

(21) *tudu* *kes* ***kumidas*** (B 25)
all that-PL food-PL
'all those foods'

(22) *kelotus* ***asesórius*** (B 100)
that.other-PL accessory-PL
'those other accessories'

There were 21 instances (4.2 %) where an NP which clearly had plural reference was not marked on any element of the NP but where plurality nevertheless could be inferred from the surrounding context (type V). In several instances, the NP had been mentioned with overt plural marking at least once in the immediately preceding text, e.g.:

(23) *tudu* ***kusa*** *nobu …* ***kusa*** *nobu* (B 82)
all thing-Ø new-Ø … thing-Ø new-Ø
'all new things ... new things'

While this configuration can be explained by the fact that the NP would still be in active memory if the examples were from authentic speech, we could not detect a systematic pattern. Note also that 18 out of the 21 occurrences come from the grammar sketch written by Ana de Paula Brito in 1887. This may reflect a change in CVC plural marking in the sense of an increase in overt plural marking: while the authenticity of Brito

(1967 [1887]) has been occasionally questioned, Quint (2005) comes to the conclusion that we are dealing with an extremely valuable and authentic source on late 19th century CVC.

There were only 11 instances (2.2 %; type VI) of outright violations of the predictions of the 4-M model, e.g.:

(24) *Kes* *ótu* ***ómi*** (B 45)
that-PL other-SG man-Ø
'those other men'

(25) *uns* ***datas*** (B 105)
INDEF-PL data-PL
'some data'

(26) *kes* *prinsipal* ***instrumentu*** *linguistiku* (VKK)
that-PL principal-Ø instrument-Ø linguistic-Ø
'those/the principal linguistic instruments'

In four out of these eleven examples, plurality was marked through *kes* occurring in a non-adjacent position to the left of the nucleus (type VII) as in examples (24) and (26). On the other hand, in

(27) ***osu*** *tudu* (B 74)
bone all
'all bones'

plurality is clearly marked albeit not in the manner predicted by the 4-M model.

7. Discussion

In this paper, we have applied the 4-M model to CVC data in order to see whether the model can be used to predict the configurations of complexity in cases of language reduction. As a matter of fact, we were able to account for the great majority of the data by applying the 4-M model to them. In a number of instances of double marking, we additionally evoked Baptista's (p.c.) observations that the factors human/animate and definiteness have the effect of triggering "redundant" nominal plural marking and related this state of affairs to a possible transformation of Portuguese singular object into CVC set nouns which the mentioned categories have resisted.

Lopes (2005) interprets her findings in the light of L2 acquisition, arguing that early system plural markers were initially retained during the massive L2 acquisition, the traces of which are still to be found in BVP.[12] The Portuguese spoken by the Tongas

12. Mello et al. (1998: 116–117) also stress the importance of language acquisition in the formation of BVP.

studied by Baxter (2004) definitely bears the imprint of L2 acquisition in earlier generations. BVP and other varieties of European languages formerly called "semi-creoles", in more recent literature "partially restructured varieties" (cf. Holm 2003: xi), have been found to owe more to L2 acquisition than so-called "full Creoles" (Winford 1997). If CVC plural marking strategies pattern quite similarly to BVP and Tonga Portuguese strategies, albeit allowing for the reinterpretation of superstrate content morphemes as Creole system morphemes (Myers-Scotton 2001: 252), one conclusion is that CVC is more of a partially restructured variety than a full Creole. Of course, "full Creole" should not be taken as an absolute construct but rather as a prototypical category (McWhorter 1998). However, further evidence on whether CVC should be classified a Creole or a restructured variety must obviously be gained from other syntactic areas than the NP and its satellites. We suggest that some diagnostic testing covering various areas found typical for at least two Ibero-Romance-based Creoles along the lines of the methodology developed by Huber and Bakker (2001) would be one possible way to proceed. As CVC presents many non-Creole features, such as partial paradigms of verb conjugation in the conjunctive, the hypothesis that CVC is rather a restructured variety than a "true" Creole no longer seems too far-fetched.

As far as the issue of complexity is concerned, we argue that in spite of being rule-governed in terms of the cognitive accessibility of different morpheme types outlined in the 4-M model, the reduction of inflectional plural marking in CVC (as well as in BVP and in Tonga Portuguese) vis-à-vis Standard Portuguese is an overall loss of morpho-syntactic complexity which can be explained in terms of L2 acquisition and language contact phenomena as well as by the shallow time-depth of the varieties in question. On the other hand, as we argue below, all instances of redundancy in plural marking imply greater complexity with regard to iconic one-to-one form-function mapping.

8. Epilogue: the cross-linguistic picture

De Groot (2005; this volume) finds that Hungarian varieties spoken outside Hungary are developing from set noun-type languages to singular object noun-type languages due to language contact with Indo-European languages. By consequence, plurality needs to be marked explicitly on nouns in the Hungarian contact varieties. As we have suggested above, the opposite appears to be occurring in Santiago CVC: the singular object nouns inherited from Portuguese may have become set nouns, with human/animate and definite nouns constituting residual categories which resist this tendency. It goes without saying that further analysis of the semantics of CVC nouns is needed to confirm this hypothesis. If this hypothesis turns out to be correct, CVC's main substrate languages should be looked at in order to determine whether we are dealing with a contact-induced change or not.

Following the metric presented by Dammel and Kürschner (this volume), the quantitative complexity of CVC plural marking appears to be fairly low as allomorphy

does not appear to play an important role.[13] For instance, there are virtually no cases of root change involved in CVC nominal formation as in Standard Portuguese.[14] Within the domain of qualitative complexity, all cases of redundant marking, albeit easier to decode, can be argued to be complex on account of Natural Morphology as the iconicity principle is violated (Wurzel 1984: 75; Kusters 2003: 53). As far as the complexity due to specific assignment principles is concerned, the semantic factors human/animate and definiteness introduce some complexity into CVC plural marking strategies which nevertheless does not reach the complexity of assignment based on grammatical gender or idiosyncratic factors found in many Germanic languages by Dammel and Kürschner (p.c.). However, as we have argued above, the main factor which conditions the assignment of overt plural markers within the CVC noun phrase is the morphological accessibility of morphemes in the formation of CVC and the 4-M model provides a framework which makes it possible to account for the specific patterns of plural marking in CVC. As Dammel and Kürschner (p.c.) restrict their inquiry to nouns for the time being, "position of the marked element within the noun phrase" is not a conditioning factor for plural assignment in their metric. Nevertheless, we consider that this position within the noun phrase and resulting accessibility of the morphemes in Creole formation is at the lower end of the complexity scale of plural allomorph assignment, perhaps right after phonological conditioning, ranked as the least complex type of conditioning in allomorph assignment. Given Dammel and Kürschner's (p.c.) different types of complexity, CVC stands very close to the lower end of the complexity scale.

References

Baptista, M. 2002. *The Syntax of Cape Verdean Creole. The Sotavento Varieties.* Amsterdam: John Benjamins.

Bartens, A. 2000. Notes on componential diffusion in the genesis of the Kabuverdianu cluster. In *Language Change and Language Contact in Pidgins and Creoles* (*Creole Language Library* 21), J. McWhorter (ed.), 35–61. Amsterdam: John Benjamins.

Baxter, A.N. 2004. The development of variable NP plural agreement in a restructured African variety of Portuguese. In *Creoles, Contact, and Language Change: Linguistic and Social Implications* (*Creole Language Library* 27), G. Escure & A. Schwegler (eds), 99–128. Amsterdam: John Benjamins.

13. As we point out in note 8 this is a shortcoming of existing discussions of CVC nominal pluralisation strategies.

14. In Standard Portuguese, we find cases like *cão* 'dog', *cães* 'dogs'. There are a few cases of "irregular synthetic plural formation" (Quint 2000: 325), e.g., *mininu* 'child', *minís* 'children'. However, *minís* should perhaps rather be described as a clipping of the coexisting regular plural *mininus* 'children'.

Brito, A. de Paula 1967 [1887]. Dialectos Crioulos-Portugueses. Apontamentos para a Gramática do Crioulo que se Fala na Ilha de Santiago de Cabo Verde. Revistas por F. Adolfo Coelho. In *Estudos Linguísticos Crioulos. Reedição de artigos publicados no Boletim da Sociedade de Geografia de Lisboa*, J. Morais-Barbosa (ed.), 329–404. Lisboa: Academia Internacional da Cultura Portuguesa.

Cardoso, E.A. 1989. *O crioulo da Ilha de S. Nicolau de Cabo Verde*. Lisboa: Instituto de Língua e Cultura Portuguesa/Praia: Instituto Cabo-Verdiano do Livro.

Castro, I. 2004. *Introdução à História do Português. Geografia da Língua. Português Antigo.* Lisboa: Edições Colibri.

Chaudenson, R. 1992. *Des Îles, des Hommes, des Langues*. Paris: L'Harmattan.

De Groot, C. 2005. The grammars of Hungarian outside Hungary from a linguistic-typological perspective. In *Hungarian Language Contact outside Hungary: Studies on Hungarian as a Minority Language,* A. Fenyvesi (ed.), 351–370. Amsterdam: John Benjamins.

Holm, J. 2003. *Languages in Contact. The Partial Restructuring of Vernaculars.* Cambridge: CUP.

Huber, M. & Bakker, P. 2001. Atlantic, Pacific, and World-wide Features in English-lexicon Contact Languages. *English World-Wide* 22(2): 157–208.

Kusters, W. 2003. *Linguistic Complexity. The Influence of Social Change on Verbal Inflection* [LOT 77]. Leiden: LOT.

Lang, J. 2002. Breve esboço da gramática do crioulo de Santiago. In *Dicionário do Crioulo da Ilha de Santiago (Cabo Verde) com equivalents de tradução em alemão e português,* J. Lang (dir.), xxii–xliv. Tübingen: Narr.

Lopes, N. de Silva 2005. Aquisição da concordância no Português: uma explicação com base na teoria dos 4M. *Papia* 15: 72–79.

McWhorter, J. 1998. Identifying the creole prototype: vindicating a typological class. *Language* 74(4): 788–818.

McWhorter, J. H. 2005. *Defining Creole.* Oxford: Oxford University Press.

Meintel, Deirdre 1975. The Creole dialect of the island of Brava. In *Miscelânea luso-africana*, Marius F. Valkhoff (ed.), 205–256. Lisboa: Junta de Investigações Científicas do Ultramar.

Mello, H. R. de, Baxter, A.N., Holm, J. & Megenney, W. 1998. O português vernáculo do Brasil. In *América negra: panorámica actual de los estudios lingüísticos sobre variedades hispanas, portuguesas y criollas,* M. Perl & A. Schwegler (eds), 71–137. Frankfurt: Vervuert.

Myers-Scotton, C. 1993. *Duelling Languages. Grammatical Structure in Codeswitching*. Oxford: Clarendon Press.

Myers-Scotton, C. 2001. Implications of abstract grammatical structure: two targets in creole formation. *Journal of Pidgin and Creole Languages* 16(2): 217–273.

Myers-Scotton, C. & Jake, J.L. 2000a. Four types of morpheme: evidence from aphasia, code-switching, and second-language acquisition. *Linguistics* 38(6): 1053–1100.

Myers-Scotton, C. & Jake, J.L. 2000b. Testing the 4-M model: an introduction. *International Journal of Bilingualism* 4: 1–8.

Quint, N. 2000. *Grammaire de la langue Cap-Verdienne. Étude descriptive et compréhensive du créole afro-portugais des Îles du Cap-Vert.* Paris: L'Harmattan.

Quint, N. 2005. *Les Apontamentois de A. de Paula Brito (1887) ou la naissance d'une tradition grammaticale capverdienne autoctone.* Paper presented at the *XIe Colloque International des Etudes Créoles*, Praia, 31.10.-7.11.2005.

Rickford, J.R. 1988. *Dimensions of a creole continuum*. Stanford: Stanford University Press.

Scherre, M. M. Pereira 1988. *Reánalise da concordância nominal em português*. Dissertação de doutorado. Rio de Janeiro: Universidade Federal do Rio de Janeiro.
Thiele, P. 1991. *Kabuverdianu: Elementaria seiner TMA-Morphosyntax im lusokreolischen Vergleich.* Bochum: Brockmeyer.
Veiga, M. 1995. *O Crioulo de Cabo Verde. Introdução à Gramática*. Praia: Instituto Caboverdiano do Livro e do Disco/Instituto Nacional da Cultura.
Veiga, M. 2000. L*e créole du Cap-Vert. Étude grammaticale descriptive et contrastive.* Paris: L'Harmattan.
Veiga, M. 2005. *Nason Global Kabuverdianu: lugar di kultura, papel di língua*. www.capeverdeancreoleinstitute.org/workshops.htm, consulted 14.8.2005.
Winford, D. 1997. On the origins of African American Vernacular English – A creolist perspective, Part I: The sociolinguistic background. *Diachronica* 14(2): 305–344.
Wurzel, W.U. 1984. *Flexionsmorphologie und Natürlichkeit. Ein Beitrag zur morphologischen Theorienbildung*. Berlin: Akademie-Verlag.

References for the corpus

APB = Brito, A. de Paula 1967 [1887]. Dialectos Crioulos-Portugueses. Apontamentos para a Gramática do Crioulo que se Fala na Ilha de Santiago de Cabo Verde. Revistas por F. Adolfo Coelho. In *Estudos Linguísticos Crioulos. Reedição de artigos publicados no Boletim da Sociedade de Geografia de Lisboa*, J. Morais-Barbosa (ed.), 329–404. Lisboa: Academia Internacional da Cultura Portuguesa.
B = Silva, T.V. da 1987. *Na bóka noti.* Praia: Instituto Caboverdiano do Livro.
BAP = Baptista, M. 2002. *The Syntax of Cape Verdean Creole. The Sotavento Varieties.* Amsterdam: John Benjamins, accompanying CD.
CNM = *Conversa num mercado*. 1979. Crioulo escrito por jovens dos bairros da Praia, Cabo Verde, 6pp.
LU = Lura 2005. *Di korpu ku alma*. Lusafrica. (CD)
NK= Silva, T.V. da 1988. *Natal y kontus*. Praia: Instituto Caboverdiano do Livro.
VDS = Veiga, M. 1982. *Diskrison Strutural di Lingua Kabuverdianu*. Praia: Institutu Kabuverdianu di Livru.
VKK = Veiga, M. 1999. *Un konbérus ba kriolu*. www.azagua.com/manuel_veiga.htm, consulted 5.8.2005.
VNG = Veiga, M. 2005. *Nason Global Kabuverdianu: lugar di kultura, papel di língua*. www.capeverdeancreoleinstitute.org/workshops.htm, consulted 14.8.2005.

Complexity and simplicity in minimal lexica

The lexicon of Chinook Jargon

Päivi Juvonen
Stockholm University

I examine the ways the minimal lexicon of a pidgin language, Chinook Jargon, gains maximal efficiency when put into use in a contemporary fictional text. The paper first describes the lexicon used from a structural point of view. It then examines the use of multifunctional lexical items in comparison to English. The results of these studies show, that 1) there is no bound morphology (neither derivational nor inflectional) in the variety studied and, 2) there is much more multifunctionality in the pidgin text than in the English texts. Finally, it is argued that the results show that the lexicon studied can indeed be described as simple and efficient.

1. Introduction

Pidginization, i.e., the creation of a restricted language variety for interethnic contact situations, fulfilling basic communication needs, dramatically minimizes not only the grammatical means speakers have at their disposal, but also the number of lexical items.[1,2] Whereas an adult speaker of for example Swedish may have a passive vocabulary of around 60,000 lexical items (The Swedish Academy Word List has 120,000 entries. See Smedberg 1906, Nagy & Anderson 1984), a speaker of a pidgin normally has only a couple of hundred words at their disposal. The total number of lexical morphemes listed in different sources for such well documented pidgin languages as Chinook Jargon, Mobilian Jargon and Lingua Franca range from around 500 to just over 2000; for each individual source, the number of lexical items is usually much lower. If, however, we disregard items listed only once or only for individual speakers, the number of lexical

1. Several people have kindly commented on this paper and in several ways improved it. I owe many thanks to Peter Bakker, Östen Dahl, Harald Hammarström, Maria Koptjevskaja Tamm, Peter Mühlhäusler and Mikael Parkvall.

2. In my use of the term, a pidgin by definition has a grammar, albeit a very limited one. Furthermore, this grammatical system strongly diverges from those of its input languages and is not straight forwardly derivable therefrom.

items in different pidgin languages in general seems to be as low as somewhere between 150 and 500 words (Peter Bakker, p.c).

Most studies on pidgin lexica so far have listed the known words and concentrated on identifying the etymological ancestry for the attested lexical items in different sources. We thus know quite a lot about the relative lexical contributions of the lexifier and the substrate languages in different pidgins. Apart from a brief overview of the observations on the organisation of pidgin lexica from a structural point of view made in different studies in Mühlhäusler (1997: chapter 5) and his observations of the semantic organisation of Tok Pisin (ibid. 1997: 156–158), the most common comment seems to be that pidgins rely heavily on polysemy and multifunctionality (see e.g., Silverstein 1972a; Holm 1989: 599). To my knowledge, however, this has not been studied in detail in any pidgin language.

The present paper focuses on the lexicon of a variety of Chinook Jargon (CJ).[3] By examining how its vocabulary is put into use in a contemporary fictional text from a structural point of view it, however, also approaches the realm of syntax, as it discusses the categorical flexibility of lexical items. This way, it aims to further our understanding of pidgin languages as such. By discussing the findings of the study in terms of simplicity and complexity it seeks to contribute to our understanding of the interplay between complexity and effectiveness of the communicative system. If a large number of elements in a system is taken to indicate complexity (e.g., Hawkins 2004: 9; Dahl 2004: 41), pidgin languages are indeed simple. In e.g., McWhorter (2001) and Dahl (2004; cf. also the papers in section one in this volume), a higher number and length of rules is also taken to indicate complexity. Therefore, this paper further examines the text in terms of lexical rules.

2. Chinook Jargon

Most pidgin languages that we know of are extinct. Many are, unfortunately, also very poorly documented: the total corpus of for example Borgarmålet, an assumed Swedish lexifier pidgin in northern Sweden, amounts to a mere five sentences (Parkvall 2001). Fortunately, other pidgins are not only well documented, but there also exist texts and even some rare tape recordings. And some pidgin languages are still in use. One of these is CJ, even though its role as an interethnic contact language is today non-existent.

CJ is a pidgin which was spoken mainly during the late 19th and early 20th century in the Pacific Northwest, along the Columbia River. It is well documented and

3. Despite its name with the epithet jargon, which could indicate an unstable variety (Holm 2000: 5–6), Chinook Jargon is here, in accordance with most researchers (e.g., Boas 1933; Thomason 1983; Harris 1994; Grant 1996a, b) considered a pidgin language with relatively high stability. For a different view, cf. Silverstein (1972a, b).

even though its glory days with some 100,000 speakers in the 1880s (Holm 1989: 597) are long gone, there are still some speakers left in British Columbia and Northwest Oregon. The main lexifier was Lower Chinook, an extinct language which had rather rich morphology (e.g., Boas 1893, Mithun 1999: 382–386, Silverstein 1972b). Besides several other Native American substrate languages (in particular Nootka), English and French have also contributed to its lexicon.

Even though the present paper focuses on the lexicon, a brief note on the grammatical system of CJ is in order.[4] Pidgin languages (typically) have no native speakers. Still, they are relatively stable linguistic systems with grammatical rules of their own. In comparison to Lower Chinook but also to other native languages in the area (and, as Silverstein (1972a, b) illustrates, also in comparison to English), CJ lacks a lot of grammatical marking. There is no trace of the three genders, the three numbers, the different possessive forms in various combinations of genders and numbers,[5] or the incorporation of subject, object or indirect object pronouns, or the locative markers of the Lower Chinook verb present in CJ; nor is there any trace of ergativity, the VS word order or the visibility distinction in the demonstrative system of Lower Chinook.

CJ is, as pidgin languages tend to be, an analytic language with practically no inflection (cf., however, Bakker 2002 on pidgin morphology and Grant 1996b, Zenk 1988 on regional and stylistic variation in CJ in this respect). The main word order is SVO, modifiers precede heads. Nominal third person subjects are repeated as free pronouns before the verb. Clausal modifiers such as negation, different kind of adverbs and interrogatives, appear clause initially. As for tenses, the Chinookan future, present and past tense can in CJ be optionally expressed by separate words. There is no number distinction in the nominal paradigm.[6] The pronominal system in CJ comprises three persons in both singular and plural. There is no distinction between subject or object forms.

---

4. The very partial enumeration of grammatical categories in Chinook is based mainly on Boas (1893; 1911), but also Mithun (1999). As Boas (1933); Grant (1996b) and Zenk (1988) have illustrated, there was quite a lot of variation between different varieties of CJ. The statements of the grammar of CJ made here are based on the standard variety described or contrastively mentioned in Boas (1933); Grant (1996a, b; 2003); Jacobs (1932); Johnson (1978); Hale (1890); Harris (1994); Holton (2004); Silverstein (1972a, b); and Thomason (1983).

5. Hale (1890) does mention that the 1st person plural inclusive is sometimes used, but this is not corroborated by other authors.

6. According to Mithun (1999: 82), North American languages typically have optional marking of number or, marking on only some nouns (mainly persons and certain animals). Jacobs (1932: 40) mentions the use of reduplication to express plural or distributive number in CJ "when the context fails to make the meaning certain". However, he also states that "such usage is very rare." In the present data there are a couple of words that seem only to have a plural or collective meaning, e.g., the word iktas translated as 'things'. See also Grant (1996b).

Possession is marked by juxtaposition or a construction of the type "man his horse". There is one main preposition, *kopa*, used chiefly in locative expressions, but a number of other items can also be used as relational words.

Apart from gender marking, which although a feature of Lower Chinook is not a very typical feature of the languages of the area, the categories mentioned above are common in the native languages of the Pacific Northwest. These languages, in particular Nootka, are further famous for blurring the boundaries of parts of speech (PoS) categories to the extent that the universality of the categories noun and verb has been questioned. Hence, it has been suggested, that some of these languages do not differentiate traditional PoS categories; rather, they operate with the categories major words vs. particles (cf. the discussion in Mithun 1999: 56–67). Major words are characterized by their ability to occur in various morphosyntactic positions, i.e., what would traditionally be called nouns, verbs and (sometimes) adjectives or adverbs, whereas particles glue the major words together into meaningful units. As these languages are rich in morphology, we can, however, distinguish between the nominal and the verbal uses.

Interestingly, categorial flexibility (or rather lack of distinct syntactic categories) does not seem to be confined to highly incorporating and morphologically rich languages such as the languages of the Pacific Northwest. On the contrary, as Gil (2005) has shown, it is also a typical feature of e.g., Riau Indonesian – an isolating language with rather poor morphology. Gil further proposes that Riau Indonesian comes very close to what he calls Isolating-Monocategorial-Associational (IMA) Language. The defining properties of IMA Language are 1) the lack of word-internal morphological structure, 2) the lack of distinct syntactic categories and 3) the lack of construction-specific rules of semantic interpretation. These properties, he argues, "represent the limiting points of maximal simplicity within each of the three domains: morphology, syntax and semantics" (ibid.: 349), and, the simplicity or complexity of natural languages (none of which is a pure IMA language) can be described in terms of how close pure IMA Language they come within the different domains.

As will soon be obvious, CJ approaches pure IMA Language in the two domains studied here: it allows words to be used in different syntactic contexts typical of different PoS categories to a high degree and, it does not have word-internal morphological structure.

3. The data

The data analysed here could be characterised as contemporary fiction in CJ. The text analysed is the story of Moola John ("Sawmill John"), written by Duane Pasco and published as a sequel in *Tenas wawa*, a newsletter about CJ (Pasco 1990–1995). Duane Pasco was born in 1932 in Seattle. According to himself (Pasco 1995) he grew up hearing CJ sporadically and had limited knowledge of it. But he set out to learn it and

started to use it in adulthood. He not only read a lot about the language, but for want of a suitable grammar, he also wrote one for contemporary use. And he started to publish texts in Chinook Jargon.[7]

English is Pasco's first language. Pidgin languages are often defined as being nobody's first language. In this regard, thus, Duane Pasco is a typical pidgin speaker. He uses standard English orthography in his writings in order to give the English reader a hint of the pronunciation, and states that the grammar is mainly based on the one found in Melville Jacobs' text collections (Jacobs 1936). Further, the vocabulary is said to be mostly from George Gibbs (1863).

If one compares the texts in Jacobs and Pasco, some differences become apparent.[8] Firstly, the texts in Jacobs (1936) are from speakers of different native languages, whereas Pasco's first language is English. Secondly, the texts are produced by speakers of different varieties of CJ (there is only one text in standard CJ, the majority of texts are in the Grand Ronde variety described in e.g., Grant 1996b). Thirdly, Pasco's text is a written text, whereas the texts in Jacobs' are transcriptions of oral stories. There are some structural differences between the texts, but as I have not thoroughly examined the grammar of neither and, (since the present paper focuses on the lexicon) I will only mention two. The short forms of personal pronouns and the short causative present in the Jacobs texts are absent in the Pasco text. A second difference between the texts is evidentiality marking. While this is not uncommon in the native languages of the area, I have not come across it in other documented CJ varieties. Still, it is not uncommon to 'know with one's eyes/ears/nose' in the Pasco text. This is rather interesting as Lower Chinook is one of the few languages in this area that does *not* mark evidentiality (de Haan 2005) and as it is explicitly stated to be uncommon in pidgins and creoles (Aikhenvald 2004: 8). Whether this is a novel development or an idiolectal use remains to be explored in still other varieties.

As regards the vocabulary in the Pasco text, with very few exceptions, all the words used were indeed found in the Gibbs' vocabulary.

The story of Moola John is, in the internet version from which the data was collected, a bilingual story. The narrative parts are mostly in English and the dialogue always in CJ, with a (close to) literal translation to English. This is what it looks like:[9]

---

7. Several people have asked me why I have chosen to analyse this particular text. The question has crossed even my mind as CJ is no longer used as an interethnic contact language. Had I had easy access to digitalized earlier texts I would most probably have chosen to analyse them instead. As it was now, however, I started with this text and found the results interesting enough to be publicly displayed. Obviously, the study presented here is a very static one. It lacks both a historical and a stylistic/variationist dimension that I hope to be able to pursue in the future.

8. For a general comment of the Jacobs' texts, see Boas (1933); for analyses of the texts see Grant (1996b); Zenk (1988).

9. In all the examples, I have followed Pasco's orthography.

Sealth:	"John, ka maika illahee?"	"John, where you country?"
John:	"Naika chako siah kahkwa konaway chee tillikum yukwa. Naika illahee siah enatai la monti wake siah hyas saltchuck ka sun chako. Yaka Boston illahee. Naika illahee nem New Hampshire."	"Me come far, same all new people here. Me country far across mountain near ocean where sun come. Him America. Me country name New Hampshire."
Sealth:	"Nu Habsha?"	"Nu Habsha?"
John:	"New Hampshire!"	"New Hampshire!"
Sealth:	"Nu Habsha!"	"Nu Habsha!"
John:	"Aha, nawitka, New Hampshire."	"Yes, correct, New Hampshire."
Sealth:	"Ikta 'Nu Habsha' kopa Chinook wawa?"	"What 'Nu Habsha' in Chinook talk?"
John:	"Yaka kahkwa chee Hampshire. Hampshire, yaka town nem kopa King Chautsch illahee."	"Him same 'new' Hampshire. Hampshire, him town name at England."
Sealth:	"Mmm, naika kumtux."	"Mmm, me understand."
John:	"Naika papa, yaka mitlite moola kopa New Hampshire."	"Me father, him have sawmill at New Hampshire."
Sealth:	"Maika papa potlatch maika moola kumtux?"	"You father give you sawmill knowledge?"
John:	"Delate. Naika pe ow mamook kopa papa moola kunsih nesaika tenas pe kwanesum mamook yahwa kimtah."	"Correct. Me and brother work at father sawmill when us small and work there always after."
Sealth:	"Maika kwan kopa Nu Habsha?"	"You happy at New Hampshire?"
John:	"Aha."	"Yes."
Sealth:	"Naika tum-tum kloshe maika chako yukwa, keschi, spose maika kwan kopa Nu Habsha, kahtah maika chako yukwah?"	"Me feel good you come here, but if you happy at New Hampshire, why you come here?"
John:	"Well, Tyee, ahnkuttie, hiyu hyas stick mitlite kopa Boston illahee, keschi delate hiyu tillikum chako yahwa. Klaska tikegh house. Moola tillikum whim stick, mamook la plash. Alta, konaway hyas stick kopet. Alta naika tikegh nanitch hyas stick yukwa."	"Well, Chief, past many big tree be at America, but very many people come there. Them want house. Sawmill people fall tree, make board. Now, all big tree finish. Now, me want look big tree here."

There were quite a few misspellings and/or orthographic variants of the same word (e.g., *ka* vs. *kah* for 'where'). I checked all these and corrected/decided upon them. The text was translated word for word, using a vocabulary compiled by the present author

by the combination of the vocabularies in Gibbs (1863) and Shaw (1909). Occasionally the word searched for was not found in the vocabulary. In these cases, the Pasco translation was consulted. This way, a parallel database with word for word alignment was created manually. Upon exclusion of proper names, words clearly meant to be in English or in a native language (as in answers to questions of the type: What is it called in English?) and, the occasional *oh*'s and *mm*'s (these latter for reasons of comparability, see below), the number of word tokens in CJ in the text is 8255, and the number of word types is 381. As a comparison: the total number of primary lexical entries in the combined Gibbs-Shaw vocabulary is 514.[10]

4. Lexical processes in a CJ text: The Story of Moola John

According to Mühlhäusler (1997: 151) pidgin lexica are "shaped by a number of constraints, including considerable pressure for one form = one meaning encoding; pressure to maximize the usefulness of a very limited lexical inventory; and the absence of processes of derivational morphology". Further, he states that in the stabilization stage (i.e., when a pre-pidgin jargon morphs into a pidgin), "circumlocution is used to express new ideas; there are a very small number of compounds at word level" (Mühlhäusler 1997: 177). Even though CJ is well beyond the stabilization stage, I will make a comment on compounding below, as it has direct consequences for the number of word tokens counted in later parts of this paper. Also, reduplication has been suggested to be common in pidgin languages. I will address these issues below.

The principle "one form equals one meaning" (Mühlhäusler 1997: 151) is perhaps best seen as a pressure to avoid synonyms, as it is common for a single lexical item in pidgin languages to encompass several related meanings (see below on polysemy). Indeed, there are few lexical items that can be considered synonymous in the Story of Moola John (MJ). One example of such a pair is *kwan* vs. *youtl*, both glossed as 'happy' by Pasco, and as 'glad' and 'proud, pleased', respectively, in the combined Gibbs-Shaw vocabulary. There are too few examples in the data to reveal any possible structural or semantic differences in their use in MJ. However, Grant (1996a: 1190) has pointed out, that there is indeed a good deal of synonymy between English and Native words in CJ. Hence, MJ seems to differ from other varieties of CJ in this respect.

An example of a systematic pattern of use of one lexical morpheme that could be argued to conform to the pressure to maximize the usefulness of a limited inventory

10. In combining the two lexica and counting the number of entries, I have only aligned the main lexical entries. Any lexemes or combinations thereof mentioned in examples and/or explanations have therefor not been counted here.

is the use of the word *mamook*. As already discussed by e.g., Silverstein (1972b), CJ has a general dynamic verb, *mamook*, meaning 'to work, to do' when used alone as in (1). But *mamook* can also be combined with other parts of speech in order to be more precise about the nature of the activity in question, as in (2, 3, 4).[11]

(1) *Naika pe ow* ***mamook*** *kopa papa moola*
1SG and brother DYANAMIC PREP father sawmill
kunsih nesaika tenas pe kwanesum Mamook yahwa kimtah.
when 1PL little and always DYANAMIC LOCDEM after
'As children, me and my brother always worked at my father's sawmill and we've been working there ever since.'

(2) Maika ***mamook*** moola ***kloshe*** ***weght?***
2SG DYNAMIC mill good again
'Can you repair the mill?'

(3) *Charlie,* ***mamook*** ***isick*** *kenkiam.*
Charlie DYNAMIC paddle right.side
'Charlie, paddle on the right side/to the right.'

(4) *Okay, tillikum, kloshe spose nesaika* ***mamook*** *skookum* ***isick*** *alta.*
Okay people good if 1PL DYNAMIC strong paddle now
'Okay people, now would be a good time to paddle all you can'

In (2)–(4) above, one could argue that there is also a possible translation that is more close to the original wording. We may also in English talk about 'making things good again' or 'using our paddles' in different ways. However, as several authors before me have noticed (see e.g., Grant 1996b; Holton 2004; Jacobs 1932; Thomason 1983; Silverstein 1972b: 614–616, who also discusses the use of the morphemes *chako* 'to come, become' and *klatawa* 'to go' as auxiliary verbs), the *mamook*-construction is a productive construction type in several CJ varieties. It (or its short form *munk*) has been described mainly as causative (Grant 1996b: 236; Jacobs 1932: 30; Mithun 1999: 589). Grant (1996b: 236) also mentions its ability to take part in lexicalization: *munk híhi* meant according to him 'to ridicule someone' and not 'to make someone laugh' (even though I have also seen several translations of it as 'to amuse someone'). Whether or not some of the uses in MJ could be seen as examples of lexicalized constructions remains, however, an unsolved issue for now. Many *mamook*-constructions in MJ can be interpreted either as main verb constructions or as causatives. Yet many others,

11. For almost all words in the lexica, several possible meanings are listed. When possible, these were summarised in terms of their grammatical category. 1SG thus stands for 'I, me, my, mine' in the combined vocabulary. The translations are my suggestions of a possible interpretation in the context where the sentences occur.

as in (2, 3, 4) above, are really neither. Rather, what *mamook* seems to function as in these constructions is as a verbalizing morpheme, i.e., it instructs you to interpret the following item as a predicate and not as an argument or, a verb rather than a noun or an adjective (see also Holton 2004: 32, who discusses also other category-changing processes in CJ). Hence, *mamook isick* means 'to paddle' and *mamook kloshe* means 'to repair'. Interestingly, as Peter Mühlhäusler (p.c.) pointed out to me, corresponding verbalizing morphemes are found in several English lexifier pidgin languages (e.g., in Samoan Plantation Pidgin English where *make horse* means 'to saddle a horse'; example from Mühlhäusler, p.c.), but also in e.g., Mobilian Jargon and Français-Tirailleur (own observations). Hence, explicit means to mark something as a predicate/verb are not confined to pidgin languages with lexifiers blurring the boundaries of traditional PoS categories (see above on Chinook). It seems to be a more general, systematic means to make the most of a small inventory of lexical morphemes.

There is no productive (bound) derivational morphology in MJ. I have found one pair of words in MJ where one could be suspected to be derived from the other by means of a derivational suffix, as their meanings are rather close. The words in question are *pahtl* glossed as 'full' (as in *pahtl yakwatin* 'full stomach') and *pahtlum* glossed as 'drunk' in the combined vocabulary and used of human beings when intoxicated with alcohol. There are quite a few words in CJ that end in *-um*: *tillikum* 'people, person', *skookum* 'strong', *sitkum* 'half, middle', many numerals etc. But there are no corresponding words without final *-um*. It does not seem to be the case that *-um* functions as a productive derivational suffix in this variety of CJ, i.e., there is no reason to postulate a derivational rule to capture the relationship between these single lexical elements.[12]

Compounding is a common means to create new lexical morphemes in many languages. Apart from structural evidence prosody is the main clue as to whether a combination of two separately analysable lexical units indeed forms a compound unit or not. As I do not have access to spoken language data, and there are few examples of e.g., expanded noun phrases, the only clue left for me is how the words or the word is spelled. As this is a shaky argument (in Swedish and Finnish there is a preference to spell compounds as one unit, whereas in English, this tendency is much weaker), I have been very reluctant to analyse any common combinations of two CJ words as compounds in MJ. In fact, the only word analysed here as a result of compounding is *saltchuck* 'ocean, sea', the combination of *salt* 'salt' and *chuck* 'water'. The reason for this is that it is in the present data more often than not spelled as one word by the native

12. However, the ending *–um* would be interesting to study in a wider perspective as it seems to be present also in some American Indian Pidgin English texts (Peter Mühlhäusler, p.c.) and as *–(j)om* also exists in Russenorsk and in the so called Solombala English of Archangel (Mikael Parkvall, p.c.).

English-speaking writer.[13] My decision has obvious consequences for any statistics on CJ vocabulary used in MJ that are presented in this paper.

Finally, a comment on reduplication, often suggested to be common in pidgin languages. As Bakker (2003) has shown, reduplication as a productive grammatical process is on the contrary rare in pidgins. As there are some morphemes in the present data that look as if they could be reduplicated forms, I will briefly comment on these. These include e.g., *moos-moos* 'buffalo', *kalakala* 'bird', *hool-hool* 'mouse' and *hee-hee* 'a laugh, to laugh' – all existing in the same form in the lexifier Lower Chinook. But there are also words like *tum-tum* 'heart, think, believe' whose etymology is uncertain. No corresponding simplex forms can, however, be found in MJ, nor have I seen these forms exemplified for other varieties of CJ: there are no words **moos*, **kala*, **hool*, **hee* or **tum*. Studying several different varieties of CJ, Grant (2003) identifies only one 'true' reduplicated form out of the 43 seemingly reduplicated forms in the pidgin (basically the variety in Gibbs 1863). This is the pair of words *pil* 'blood' and *pilpil* 'red'. However, in the present data the form *pil* is used for both these meanings, according to Pasco's translation. Hence, there does not seem to be any reason to state that reduplication is a productive lexical process in MJ, nor was a feature of CJ lexicon in general.

To summarize, the variety of CJ studied here does conform to the constraints shaping pidgin lexica proposed by Mühlhäusler (1997: 151). There are few synonyms and no productive (bound) derivational morphology. Also, a systematic means to make maximal use of a small inventory was exemplified. The present data neither support nor disconfirm that there be a small number of lexical compounds, but the issue was addressed as it has direct bearing on the statistics presented shortly.

5. Recycling meaning in a CJ text: the multifunctionality of lexical items

Pidgin languages are often said to rely heavily on polysemy and multifunctionality. Indeed, the combined Gibbs-Shaw vocabulary includes numerous examples of words listed as having several related meanings. As an example, Johnson (1978: 222) pointed out that the CJ word *tilikam* (spelled *tillikum* in the present data) has for example been variably analyzed as 'man', 'people', 'person', 'Indians', 'natives', 'relatives', 'kindred', 'non-chiefs', 'same tribe', 'same band', and 'friend'. All these semantically related meanings seem to belong to the class of nouns. However, there are also semantically

13. There is one general exception to my principle to count two orthographic words as separate lexical morphemes, called agglutination in creolistics. Words of French origin have in CJ, as they have in many French-lexicon pidgins and creoles, the original definite article hanging along, as in *la monti* 'mountain' (< *la montagne* 'the mountain') or *la push* 'mouth' (< *la bouche* 'the mouth'). These remnants of articles do, however, not carry along their original meaning and are, thus not counted as separate words, but as parts of monomorphemic nouns. For more examples of agglutination in pidgins, cf. Mühlhäusler (1997: 154–155).

different but related meanings connected to single lexical morphemes, where a word class boundary has been crossed. Consider the English sentences (5, 6) (taken from P.G. Wodehouse's 1911 novel *The best sauce*).

(5) He fell in *love* with Eve at sight, and if, at the end of the first day, there was anyone in the house who was not aware of it, it was only Hildebrand, aged six.
(6) Perhaps I am the human parsnip, and you will have to learn to *love* me.

In (5), the word *love* appears in a typical syntactic position for a noun, in (6) in a typical position for a verb. Obviously, these uses are related to each other, but they appear in different syntactic environments. This subset of polysemous lexical items, i.e., items that can appear in positions typical of different PoS categories and that can be identified on syntactic grounds, is what is meant by multifunctionality in the present paper (see Enfield 2006 on heterosemy; cf. also Mühlhäusler 1997 and references therein). As there are plenty of examples of polysemy in pidgin lexica in the literature (e.g., Gibbs 1863; Johnson 1978; Shaw 1909), I will in the following concentrate on the much less studied matter of multifunctionality.

As already illustrated, there are multifunctional lexical morphemes in English. These are not uncommon in English, at least not if compared to other Germanic languages, not to mention more agglutinating languages such as Finnish. Anecdotally, when learning English, you are told that there are common words that can function both as verbs and nouns such as *love*, *talk*, *walk*, *stone* etc. When learning Swedish or German, nobody bothers to state this (even though they exist). However, in terms of how common multifunctionality is in a language, English does not seem to be anywhere in the vicinity of the isolating languages of South East Asia (Enfield 2006 and references therein; Gil 2005 and references therein), the languages in the Pacific Northwest (Mithun 1999, see also the discussion in section 1 above), creoles (Lefebvre 2001) and – as will be argued here with reference to the present data – pidgins.

One of the things that are relatively stable in a pidgin language is the order of lexical items, even though there is some variation in this respect as well (Mühlhäusler 1997: 144–146 and references therein). I have tried to be open with regard to the data at hand when identifying different syntactic environments, but there is of course always the possibility that I have been influenced by my own linguistic background in the analyses. This is why I will give numerous examples of lexical items that I have counted as multifunctional. As previously briefly discussed (section 1), the languages in the Pacific Northwest are famous of being reluctant to an analysis in terms of (Eurocentric?) PoS categories. Despite this, and despite the fact that almost all writers in one way or another acknowledge that multifunctionality (as defined here) is a common feature of CJ, terms like noun, verb, adjective etc are commonly used in describing the language. I will continue this tradition, as I am not convinced that this is such a serious mistake when it comes to this particular variety of CJ (as there are words with less flexible syntactic behaviour than those discussed at present) and, as trying to do without these terms would require a lot more space than I have at my disposal.

CJ does not have an overt copula. Thus, we have the locative construction *John, ka maika illahee*? 'John, where your country?' and the equative construction *Ikta 'Nu Habsha' kopa Chinook wawa*? 'What 'Nu Habsha' in Chinook Jargon?' in the extract from the story above. Besides the zero-copula construction, there are two main clause schemas in MJ, one for 1st and 2nd person subjects (nominal and pronominal), but also 3rd person pronominal subjects and, another for 3rd person nominal subjects:

(ADVERB*)	SUBJECT	VERB*	(OBJECT-NP)	(PP/ADVERB*)

3rd person NOMINAL SUBJECT	(ADVERB*)	3rd person SUBJECT PRO	VERB*	(OBJECT-NP)	(PP/ADVERB*)

Even though these schemas indeed seem to cover the structure of most of the clauses in MJ, there is still some variation as regards the position of adverbs. The usual internal structure of an NP with a nominal head can be roughly schematized as:

(QUANTIFIER/PRO)	(ADJECTIVE)	(NOUN(PRO))	NOUN	(PP/ADVERB)

where the optional first noun, if not followed by a pronoun, is according to Holton (2004: 36) always interpreted as an adjective. In some cases this might be so (*Boston tillikum* 'American people'), but I am not convinced that this is always the case (*moola engine*, literally 'sawmill engine' is a case in point). The noun-pronoun-noun construction has a possessive reading (*tyee yaka klootchman*, literally 'chief his woman').

There are words in MJ that appear both in the environment immediately following the subject or the resumptive 3rd person subject pronoun (in this case without intervening adverbs) as in (7) and, in the environment following a preposition (and in this case also a modifier) as in (8). These words, thus, function as both verbs and nouns.

(7) *Tyee, klonas kloshe spose maika* ***wawa*** *okoke ekahnum?*
chief maybe good if 2SG tell DEM story
'Chief, maybe it would be nice if you told this story?'

(8) *Ikta kahkwa skookum man kopa Chinook* ***wawa***
3SG like strong man PREP Chinook talk
'His name means 'a strong man' in Chinook Jargon.'

Besides the word *wawa*, I have analysed the following words as belonging to this category: *camp* 'to camp, a camp'; *chanti* 'to sing, a singer'; *chukkin* 'to kick, a kick'; *coolie* 'to run, a river'; *hee-hee* 'to laugh, laughter'; *hullel* 'to shake, a shake'; *huy-huy* 'to trade, trading'; *isick* 'to paddle, a paddle' (appears also with *mamook* as in (3)); *mukamuk* 'to eat, food'; *whim* 'to fall, to drop, a log'; *pait* 'to fight, a fight'; *potlatch* 'to give, a party (thanksgiving)'; *smoke* 'to smoke, a smoke'; *snass* 'to rain, a rain'. This is the largest category found in these data in terms of types of words.

There are words that are used as both verbs (immediately after the subject) and as adverbs (before the subject) as in (9, 10).

(9) *Wind* ***elip*** *poh pe salt chuck chako peshak*
wind begin blow and sea become bad
'The wind started to blow and the sea became enraged.'

(10) *Ahnkuttie,* ***elip*** *Boston Tillikum chako, nesaika*
formerly before American People came 1PL
tillikum mitlite yahwa.
people live LOCDEM
'In the past, before the Americans arrived, our people lived there.'

Elip is in MJ also used to mean 'first'. Besides the word *elip*, the words *kilapai* 'to return, back'; and *kopet* 'to finish, to stop, just, only', appear in similar constructions.

In (3), (repeated here) the word *kenkiam* is used in a position where one can find adverbs like 'quickly', but also PPs like 'to the right side'.

(3) *Charlie mamook isick* ***kenkiam***
Charlie DYNAMIC paddle right.side
'Charlie, paddle on the right side/to the right.'

Regardless of whether we analyse *kenkiam* in (3) as an adverb or something else, it contrasts with the position it has in (11), where *kenkiam* appears before a noun as a modifier.

(11) *Hamitchou kow klone Klapite pahtl hykwah kopa yaka*
Hamitchou tie three String full dentalium PREP 3SG
kenkiam *okchoke pe klone klapite phatl hykwah kopa*
right shoulder and three string full dentalium PREP
yaka huloima okchoke.
3SG other shoulder
'Hamitchou tied three strings full of dentalium on his right shoulder and three strings full of dentalium on his other shoulder.'

There are also words that are used both in the position normally associated with prepositions (12) and as (free) clausal adverbs (13).

(12) *Hamitchou nanitch* ***kimtah*** *yaka pe kumtux*
Hamitchou look behind 3SG and know
kopa eye ikt hyas nenamooks.
PREP eye one big otter
'Hamitchou looked behind himself and saw an enormous otter with his own eyes.'

(13) ***Kimtah,*** *nesaika kloshe sihks.*
after 1PL good friend
'Afterwards we became good friends.'

As usual when analysing empirical data, many examples are difficult to analyse. E.g., in (14), the word *ahnkuttie* is used twice; first clause initially, i.e., in the adverb position, but the second time in a position for a modifier within a noun phrase. I have analysed these as two different syntactic positions, even though I dare not even suggest a PoS category label for the second *ahnkuttie*!

(14) *Siah* ***ahnkuttie****, naika* ***ahnkuttie*** *tillikum mitlite kopa Dungeness.*
far before.now 1SG before.now people live PREP Dungeness
'Long time ago, my ancestors lived in Dungeness.'

Besides the patterns of multifunctionality in (3, 7–14), there are many more, which space does not allow me to demonstrate. Some of the (perhaps) less exciting but common patterns are that wh-words function also as conjunctions, adjectives appear in positions for (free) adverbs, and the numeral *ikt* 'one' appears both as such and in an indefinite article-like fashion.

As already mentioned, it has been claimed that multifunctionality is a characteristic, i.e., a frequent feature of pidgin languages. In order to see how frequently words where used in different syntactic positions in MJ, I first compiled a simple frequency list of the words used. After that, I checked the syntactic environment of all words that appeared more than once in the text. Quite a few appeared in syntactic environments characterized above as typically belonging to different PoS categories. After that, I compiled frequency lists of collocations, i.e., two words occurring together in the text (bigrams) in order to check that I had not missed any collocations/syntactic environments when performing the analysis manually.

There are plenty of examples of uses of words in different syntactic environments that I have not counted as examples of multifunctionality. For example, I have not counted any of the words in constructions such as *tikegh iskum kumtux* in (15) below as multifunctional, neither have I counted use as auxiliary vs. main verb as multifunctionality, nor the use of one and the same pronoun in subject, object and possessive constructions, which occurs regularly – simply because they can still be considered to retain whatever category they belong to.

(15) *Tyee, naika* ***tikegh iskum kumtux sit*** *kahkwa siwash tillikum!*
chief 1SG want get know sit like Indian people
'Chief, I want to learn how to sit like a native!'

Also, as hinted at above in connection to example (14), it is sometimes hard to be exact about the PoS category of words. This is a common problem when studying cases of incipient grammaticalization, as for example when the numeral *ikt* appears in more article-like functions. For these reasons, and for reasons of comparability with English, only cases where at least one of the uses of a multifunctional word was identified as that of a typical noun, verb, adjective or an adverb (i.e., major, open class PoS categories) are accounted for here. Table 1 display to what extent multifunctionality in this narrow sense is present in the story of Moola John.

Table 1. Cross-category multifunctional lexical items in the CJ text Moola John.

	Multifunctional	Monofunctional	Total
Word forms	42	339	381
%	11%	89%	100%
Word tokens	1073	7182	8255
%	13%	87%	100%

In order to see, whether these figures were high or low in comparison to a text in another language, I calculated the rate of multifunctionality also in original English text. Optimally, I would have wanted to calculate this in an English version of the Story of Moola John. However, there was no standard English translation available: Pasco's translation was not a standard English version of the story but, rather, a literal translation. And, upon myself trying to translate the text into English, I soon realized that this was not feasible: apart from the fact that I am not a native English speaker, the nature of CJ proved this impossible. It very soon became apparent, that more often than not, a CJ meaning could be translated in several partly different ways. E.g., should the CJ *Siwash tillikum* be translated as 'Indians', as 'native people' or as 'natives'; *klatawa* sometimes as 'go', sometimes as 'leave', sometimes as 'travel' etc? Whatever I would choose, it would have had an obvious effect on the results of a later multifunctionality count. Because of this risk of distorting the results I calculated the rate of multifunctionality in two English short stories instead. The calculations were made in the same fashion as for the MJ text. The first text is a short story called *Second hand rows* (Coleman 1993). However, as the total number of words in this short story is rather low, and as the rate of multifunctional items could depend on the difference in text length, I also analysed a longer English text: a short story called *The best sauce* (Wodehouse 1911). The stories were chosen from a digital source because they 1) included some dialogue (MJ is basically a dialogic text) and 2) the total number of words in them approached the number of words in the MJ data. Table 2 and 3 give the comparable figures.

Table 2. Cross-category multifunctional items in English, *Second hand rows*.

	Multifunctional	Monofunctional	Total
Word forms	4	757	761
%	0.5%	99.5%	100%
Word tokens	15	2282	2297
%	<1%	>99%	100%

As can be seen, there are only four word forms that function in this way in this text. These are *end*, *run*, *back* and *break*.

Table 3. Cross-category multifunctional items in English, The best sauce.

	Multifunctional	Monofunctional	Total
Word forms	13	1357	1370
%	1%	99%	100%
Word tokens	50	4858	4908
%	1%	99%	100%

The word forms found to be used across PoS categories in this story were *last*, *look*, *stay*, *love*, *play*, *back*, *clear*, *rest*, *sort*, *count*, *laugh*, *pack* and *return*. Examples (5, 6) above are taken from this short story. As can be seen in table 3, both the number of word types and word tokens that are used in different syntactic environments increase slightly in comparison to those accounted for in table 2, but the differences against the analysis of the CJ text remain. Hence, whereas the proportion of multifunctional words (as defined and counted here) in an English fictional text seems to land around 1% in terms of both word types and word tokens, the proportion of multifunctional words in the CJ text studied is 13% in terms of word types, and 11% in terms of word tokens. CJ, thus, uses a much higher proportion of multifunctional lexical items than English. In light of the present data, multifunctionality can indeed be said to be a characteristic feature of at least the one pidgin lexicon examined here. I would expect other CJ varieties as well as other pidgin languages to behave similarly, although this remains to be proven.

6. The lexicon of Chinook Jargon – simple and efficient

As Silverstein (1972b: 614) notes with reference to CJ, despite ". . . lack of complexity in syntax, which we can characterize as structurally shallow and discretely linear, there remains the possibility that the lexicon is characterized by complexity." The present study has examined how the small vocabulary of CJ is put into use in a contemporary fictional text from a structural point of view. In this section, I will briefly discuss the findings heretofore in terms of structural complexity. Finally, I will address the question of the relation of system complexity and its efficiency.

If we adopt the distinction made by Dahl (2004) between resources and regulations in discussing complexity (see also Dahl in this volume), the number of words in a language clearly belongs to the realm of resources. Compared to fully fledged languages (creoles or non-creoles), pidgin languages have a lot less of lexical resources and are thus in this respect less complex. However, as Dahl (2004: 41) points out, even though the distinction between lexicon and grammar is reminiscent of that between resources and regulations, it does not completely coincide with it. Firstly, a lexicon may be regulated, i.e., there may be rules in a lexicon. And secondly, in the information theoretic

spirit of Dahl, we might need longer or shorter descriptions of the lexical items there are, i.e., the lexical items themselves may be more or less complex.

There were no bound morphological derivational rules in the data studied and no lexical reduplication. Also in this respect, the lexicon of CJ is simple. There was, however, a systematic means to create active predicates by means of a construction, where the verb *mamook* combines with other verbs, nouns and adjectives. Irrespective of whether we look upon it as a grammatical or a lexical construction, it can be described as "shallow and discretely linear" (cf. the citation on CJ grammar from Silverstein above). We can analyse the parts of the construction and compose the new meaning directly from the parts. Hence, in regard to structural complexity in terms of different types of rules (compositional vs. disintegrative), the lexicon of CJ remains simple.

From another angle, the lexicon of CJ was described above as a multitude of structurally vague multifunctional items. Obviously then, if we are to describe the use of a single multifunctional lexical item, the description needs to be longer than when describing items with one basic meaning (if we for the time being disregard the multitude of functions basically non-multifunctional items can be put into in interaction). In this respect, from an information theoretic and thus "at least in principle "objective" angle" (Dahl 2004: 39), there is *some* complexity in the lexicon of CJ (see also the discussion of "choice structure" in Dahl 2004: 46–50).

Hence, the lexicon of CJ (in light of the data studied) can indeed be said to be simple from an information theoretic point of view. And, if Silverstein's remark of the simplicity of its grammar is correct (which I indeed take it to be), we must conclude that CJ, as a system, is a simple one. Pidgin languages are makeshift varieties fulfilling basic communicative needs. They are usually used in limited domains. They (typically) have no native speakers. In Dahl's (2004) terms, they are characterized by small resources and a minimum of regulations. From the point of view of the people who use them, they thus need to make maximum use of a minimal lexicon (Mühlhäusler 1997: 158); they need to be effective as means of communication. This is exactly what they are. According to Hawkins (2004: 9; see also Gil 2005):

> Efficiency is increased, first, by minimizing the domains (i.e., the sequences of linguistic forms and their conventionally associated properties) within which certain properties are assigned. It is increased, secondly, by minimizing the linguistic forms (phonemes, morphemes, etc.) that are to be processed and by reducing their conventionally associated properties, maximizing in the process the role of contextual information (broadly construed), including frequency effects and various inferences. Third, efficiency is increased by selecting and arranging linguistic forms so as to provide the earliest possible access to as much of the ultimate syntactic and semantic representation as possible.

This, I believe, is how it works in the Chinook Jargon text studied here. Even though I expect it, it remains to be seen, of course, whether this holds also for other varieties of CJ and other pidgin languages.

Abbreviations

1SG	first person singular pronoun
2SG	second person singular pronoun
1PL	first person plural pronoun
DEM	demonstrative pronoun
DYNAMIC	the dynamic verb *mamook*
LOCDEM	demonstrative locative adverb
NP	noun phrase
PP	preposition phrase
PREP	preposition

References

Aikhenvald, A.Y. 2004. *Evidentiality.* New York: Oxford University Press.

Bakker, P. 2002. Pidgin inflectional morphology and its implications for creole morphology. In *Yearbook of Morphology*, P.A. Doevendans, G. Booij & J. van Merle (eds), 3–33. Dordrecht: Kluwer Acadamic Publishers.

Bakker, P. 2003. The absence of reduplication in Pidgins. In *Twice as meaningful. Reduplication in Pidgins, Creoles and Other Contact Languages,* S. Kouwenberg (ed.), 37–46. London: University of Westminster Press.

Boas, F. 1893. Notes on the Chinook language. *American Anthropologist* 6(1): 55–64.

Boas, F. 1911. Chinook. In *Handbook of American Indian Languages*, F. Boas (ed.), 559–677. Washington: Government Print Office. Smithsonian Institution, Bureau of American Ethnology.

Boas, F. 1933. Note on the Chinook Jargon. *Language* 9: 208–213.

Coleman, L. 1993. Second hand rows. *Whole Earth Review* 78.

Dahl, Ö. 2004. *The Growth and Maintenance of Linguistic Complexity*. Philadephia, PA, USA: John Benjamins Publishing Company.

Enfield, N.J. 2006. Heterosemy and the grammar-lexicon trade-off. In *Catching Language: The Standing Challenge of Grammar Writing,* F. Ameka, A. Dench & N. Evans (eds), 1–24. Berlin: Mouton de Gruyter.

Gibbs, G. 1863. *A Dictionary of the Chinook Jargon, or Trade Language of Oregon*. Washington: Smithsonian Institution.

Gil, D. 2005. Isolating-Monocategorial-Associational Language. In *Handbook of Categorization in Cognitive Science*, H. Cohen & C. Lefebvre (eds), 349–379. Oxford: Elsevier.

Grant, A. 1996a. Chinook Jargon and its distribution in the Pacific Northwest and beyond. In *Atlas of Languages of Inter-cultural Communication in the Pacific, Asia and the Americas,* S.A. Wurm, P. Mühlhäusler & D.T. Tryon (eds), 1184–1208. Berlin & New York: Mouton de Gryuter.

Grant, A. 1996b. Functional categories in Grand Ronde Chinook Jargon. In *Changing Meanings, Changing Functions,* P. Baker & A. Syea (eds). London: University of Westminster Press.

Grant, A. 2003. Reduplication in Chinook Jargon. In *Twice as Meaningful. Reduplication in Pidgins, Creoles and Other Contact Languages,* S. Kouwenberg (ed.), 319–322. London: University of Westminster Press.

Haan, F. de 2005. Coding of Evidentiality. In *The World Atlas of Language Structures,* M. Haspelmath, M. Dryer, D.Gil & B. Comrie (eds), 318–321. Oxford: Oxford University Press.

Hale, H. 1890. *An International Idiom. A Manual of the Oregon Trade Language, or "Chinook Jargon".* London: Whittaker & Co.

Harris, B.P. 1994. Chinook Jargon: Arguments for a pre-contact origin. *Pacific Coast Philology* 29(1): 28–36.

Hawkins, J.A. 2004. *Efficiency and Complexity in Grammars*. Oxford, New York: Oxford University Press.

Holm, J. 1989. *Pidgins and Creoles.* Vol II. New York: Cambridge University Press.

Holm, J. 2000. *Introduction to Pidgins and Creoles.* West Nyack, New York: Cambridge University Press.

Holton, J. 2004. *Chinook Jargon: The Hidden Language of the Pacifc Northwest.* San Leandro, CA: Wawa Press.

Jacobs, M. 1932. Notes on the Structure of Chinook Jargon. *Language* 8(1): 27–50.

Jacobs, M. 1936. Texts in Chinook Jargon. *University of Washington Publications in Anthropology* 7(1): 1–27.

Johnson, S.V. 1978. Chinook Jargon: A Computer Assisted Analysis of Variation in an American Indian Pidgin. PhD Dissertation, University of Kansas.

Lefebvre, C. 2001. Multifunctionality and the concept of lexical entry. *Journal of Pidgin and Creole Languages* 16(1): 107–145.

McWhorter, J. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology* 5-2/3: 125–166.

Mithun, M. 1999. *The Languages of Native North America.* Cambridge: Cambridge University Press.

Mühlhäusler, P. 1997. *Pidgin and Creole Linguistics.* Expanded and revised edition. London: University of Westminster Press.

Nagy, W.E. & Anderson, R.C. 1984. How many words are there in printed School English? *Reading Research Quarterly* 19(3): 304–330.

Parkvall, M. 2001. Borgarmålet. Paper presented at The Westminster Pidgin Workshop, London, April 2001.

Pasco, D. 1995. Epilogue to the *Tenas Wawa*. Downloaded from http://tenaswawa.home.att.net/jn14.htm 2006-03-28.

Pasco, D. 1990–1995. Moola John. Downloaded from http://tenaswawa.home.att.net/jn14.htm 2005-04-14.

Powell, J.V. 1990. Chinook Jargon vocabulary and the lexicographers. *International Journal of American Linguistics* 56(1): 134–151.

Shaw, G.C. 1909. *The Chinook Jargon and How to Use it: a Complete and Exhaustive Lexicon of the Oldest Trade Language of the American Continent*. Seattle: Rainier Printing Co.

Silverstein, M. 1972a. Chinook Jargon: Language contact and the problem of multi-level generative systems, I. *Language* 48: 378–406.

Silverstein, M. 1972b. Chinook Jargon: Language contact and the problem of multi-level generative systems, II. *Language* 48: 596–625.

Smedberg, A. 1906. Några tankar rörande allmogespråkets ordförråd. *Svenska landsmål ock svenskt folkliv* 11: 1–28.

Thomason, S. G. 1983. Chinook Jargon in areal and historical context. *Language* 59: 820–870.

Wodehouse, P.G. 1911. The best sauce. *The Strand Magazine.*

Zenk, H. 1988. Chinook Jargon in the speech economy of Grande Ronde Reservation, Oregon: an ethnography-of-speaking approach to an historical case of creolization in process. *International Journal of the Sociology of Language* 71: 107–124.

Index of languages

A
Abkhaz 274
Acehnese 173–4
Acoma 274
Adang 70, 78–9
Adyghe 266
Afrikaans 244, 247, 249, 253, 256, 258
Ainu 274
Akha 183–4
Akkadian VII
Alamblak 274
Amele 274
American Indian Pidgin English 329
Anêm 231–2
Annobonese 278, 298
Apurinã 275
Arabic 12, 19, 96, 126, 168–9, 274, 278, 280, 298
  Classical, 12
  Egyptian, 274
  Modern Standard, 168
  Kinubi, *see* Kinubi
Arapesh 70, 77, 79, 274
Archi 109
Armenian, Eastern, 275
Asiluluan 180
Asmat 178, 274
Assamese 299
Australian Creole 278
Awa Pit 275

B
Babungo 70, 79, 82, 84
Bagirmi 274
Baham 179
Bahasa Nusa Laut 180
Baikenu 177
Balanta 179
Banda-Linda 185
Barasano 275
Basque 273–4
Batumeran 180
Beja 274
Berber, Middle Atlas, 70, 80, 275
Berbice Dutch Creole 69, 71, 79
Bima 186
Bislama 118, 120–1, 278, 299
Bonfian 180
Borgarmålet 322
Brahui 275
Breaux Bridge 298
Breton 294, 296
Broken, Torres Str., 299
Bunak 178
Burmese 274
Burushaski 274

C
Cahuilla 275
Cameroon Pidgin 57–8, 299
Camus 290–1, 297
Canela-Krahô 275
Cantonese 120–1
Cantonese Pidgin 299
Cape Verdean Creole Portuguese, *see* Kabuverdianu
Carriacou 299
Catalan 50, 52–3
Cayuvava 275
Chamorro 274
Chinese XI, 8–9, 124, 133–4, 167–9, 179, 185, 187, 266, 269
  Cantonese, *see* Cantonese
  Classical, 266, 269
  Mandarin, *see* Mandarin
Chinook 266, 299, 323–5, 329–30
  Lower, 323–5, 330
Chinook Jargon XIII, 278, 299, 321–2, 325, 332, 336–7
Choctaw 299
Chukchi 275
Coastal New Guinea Pidgin 299
Comanche 274
Coos 275
Cora 71, 78–9
Cree, Plains, 70, 80, 274
Crioulo Guiné 298
Croatian 50, 53

D
Daga 70, 79, 275
Dalecarlian 161
Dani, Lower Grand Valley, 275
Danish 244, 247, 249, 251, 254–6, 258–9, 270
Dawanese 177–9
Degema 185
Dení 156
Diola-Fogny 275
Diyari 70, 79
Dominican 278
Dutch 94, 97, 99–101, 203, 209, 244–5, 247, 249, 251, 253, 255–6, 258–9, 278, 299
Dyirbal 274

E
English X–XIII, 26, 29, 43–4, 50–4, 56–61, 72, 91, 94, 96–102, 109, 114–5, 118, 120–2, 124–8, 133, 135, 140–3, 145–6, 153–8, 161, 167–72, 181, 187, 195, 199, 207, 218–9, 230–1, 243–4, 247, 249, 252–3, 255–6, 258–60, 266–71, 275, 277–8, 289, 296, 298–9, 308, 321, 323, 325, 327–31, 334–6
  Modern, 54, 168–9, 195
  Old, 54, 169, 195

Epena Pedee 275
Erromangan 295
Eskimo Trade Jargon 299
Esperanto 268–9
Estonian 97, 221
Evenki 274
Ewe 275

F
Fanakalo 170, 187, 278, 298–9
Faroese xii, 243–4, 247–50, 252, 254–6, 258, 260
Fataluku 178–9
Fijian 274
Finnish 67, 94, 97, 99–101, 104, 145, 219, 221, 269, 275, 329, 331
Français-Tirailleur 329
French 94, 97, 99–101, 110–2, 123–4, 134, 138–9, 153, 266–7, 269, 271, 274, 277–8, 298, 323, 330
Frisian 244, 247, 253, 256, 258
  West, 244, 247

G
Galoli 177, 181
Georgian 70, 79, 274
German 3, 9, 46–7, 50–3, 101, 153, 161, 168, 183, 209–10, 219, 244, 247, 249–51, 253–6, 258, 260, 271, 274, 331
Gooniyandi 70, 78–9, 274, 289
Grebo 274
Greek 101, 135–6, 139–40, 142–3, 146, 157, 274, 292
  Modern, 101, 135
Greenlandic, West, 70, 77, 80, 270, 275
Guadeloupean Creole 278
Guajiro 294
Guaraní 71, 74, 153, 155, 275
Guinea-Bissau Creole 179, 278, 298, 300

H
Haida 275
Haitian Creole 101, 104, 146, 277–8, 298
Harukuan 180
Hausa 275
Hdi 34
Hebrew 97, 120–2, 275
Highlands New Guinea Pidgin 299
Hindi 274
Hiri Motu 299
Hitunese 180
Hixkaryana 71, 80, 275
Hmong xi, 70, 110, 133–8, 142–5, 147, 149, 275
  Njua 275
Hungarian xii, 70, 72, 80, 98, 101, 104, 191–2, 194–5, 198, 200–13, 219, 269, 274, 317
Hunzib 274

I
Iban 172
Icelandic 111, 161, 243–4, 247, 249–52, 254, 256, 258
Igbo 274
Iha 175, 179
Ika 71, 78, 80, 275
Imonda 275
Indonesian xi, 70, 80, 101, 114–5, 120–1, 124–8, 134, 145, 167–8, 170–5, 178, 183–4, 186–7, 269, 275, 288, 324
  Riau, 114–5, 124–8, 134, 145, 167, 171–5, 186, 324
  Standard, 168, 171–4, 186
Ingush 274
Inuit, MacKenzie, 299
Iranian 168
Iraqw 275
Italian 50–1, 53, 278

J
Jakaltek 274
Jamaican Creole 278, 299
Japanese 91, 269, 275
Jaqaru 71, 80
Jarawara 37
Javanese 110, 145, 172
Ju|'hoan 274

K
Kabardian 8
Kabuverdianu xii, 298, 305–7, 309, 311–8
Kambera 176, 178
Kannada 70, 74–5, 77–80, 274
Kanuri 274
Kap-Verde Creole, *see* Kabuverdianu
Kare 293
Karipuna, Amapá, 298
Karok 35, 274
Kayardild 274
Kayeli 180
Kemant 34–5
Keo 181–8
Ket 275, 296
Kewa 274
Khalkha 274
Khasi 275
Khmer 173, 275
Khoekhoe 70, 80, 274
Kikongo 170, 174, 294
Kilivila 275
Kimbundu 310
Kinubi 12, 169, 278, 298
Kiowa 34–5, 275
Kisi 70, 78, 80
Kituba 170, 187
Koasati 71, 75, 80, 274
Kobon 275
Kombai 70, 80
Korean 274
Korku 70, 78–80
Koyraboro Senni 35, 275
Krio 278, 299
Krongo 275
Kua 266
Kunama 275
Kuot xii, 70, 78, 80, 217–34, 237, 240, 242
Kutenai 275
Kwa 135
Kwazá 24

L
Ladakhi 273–4
Lahu 275
Lakhota 275
Lango 275
Lao 111, 124, 134–5
Larikean 126, 180
Latin 3, 8, 100, 110–2, 123–4, 169, 267–8, 277
Latvian 274
Lavukaleve 70, 79–80, 275
Lezgian 274
Lingala 170
Lingua Franca xiii, 278, 321

Lusi 231
Lutuami 293
Luvale 275
Luxembourgish 244, 247, 249–54, 256, 258, 260

M
Macedonian 3, 268
Madak 217–20, 222–6, 228–30
Maduresse 172
Maipure 295
Makalero 178
Makasai 178
Malagasy 274, 295
Mambai, Ainaro 177
Mandarin xi, 110, 124, 133–8, 143–4, 147, 149, 167–8, 179, 187, 269, 274
Mangarrayi 275
Manjaku 179
Maori 94, 97, 99–100, 274
Mapudungun 275
Maricopa 70, 79–81, 274
Martinican Creole 278
Martuthunira 274
Masarete 180
Maung 274
Mauritian Creole 278, 298
Maybrat 70, 80, 275
Meithei 274
Mien 70, 80
Minangkabau 118, 120–4, 172–3
Miwok, Southern Sierra, 80, 275
Mixtec, Chalcatongo, 275
Mobilian Jargon xiii, 299, 321, 329
Mon-Khmer 173
Mor 179
Muna 180
Mundari 274
Murinypata 295–6

N
Naga Pidgin 299, 300
Nahuatl, Tetelcingo, 275
Nalik 217–20, 222–30
Nasioi 34
Ndyuka 8, 275–9, 281, 299
Negerhollands 278, 299
Nenets 275
Nez Perce 274
Ngadha 181–3, 185–8
Ngala 170
Ngbandi 299
Ngiti 70, 80
Ngiyambaa 274
Nialan 180
Nigerian Pidgin 278, 299
Nivkh 274
Nootka 323–4
Norse, Old, 19
Norwegian 14, 161, 278
Notsi 217–20, 222–30, 236, 239
Nubi Creole Arabic, *see* Kinubi
Nubian, Dongolese, 70, 80
Nunggubuyu 275
Nuuchahnulth 71, 80, 82, 84
Nyokon 293, 297

O
Oneida 274
Oromo, Harar, 275
Osage 71, 80, 84
Otomí, Mezquital, 274

P
Paiwan 275
Palenquero 174, 278
Panjabi 291, 294, 296, 300
Papia Kristang 278, 298
Papiamentu 120–2, 278, 299–300
Passamaquoddy-Maliseet 274
Paumarí 275
Persian xi, 167–8, 187, 275
Pig Latin 268
Pirahã 71, 80, 275
Pointe Coupée Creole 298
Polish 133, 135–6, 140–3, 146
Polynesian 173, 175, 178
Portuguese xiii, 50, 52–3, 101, 179, 181, 278, 298, 305–6, 308–12, 315–8
  Brazilian, xiii, 305, 309–11, 316–7
  European, 309–10
  Tonga, 317
Principense 278, 298

Q
Qiang 70, 76, 78, 80
Quechua x, 3–4, 6–7, 13, 15–20, 71, 80, 274
  I, 16
  II, 16–7
  IIb, 16–7
  IIc, 16
  Argentinean, 13, 16, 18–1
  Bolivian, 6–7, 13
  Cuzco, 6–7, 13, 16–8, 20
  Ecuadorian, 13, 16–9
  Imbabura, 274
  Inca, 16

R
Rama 275
Rapanui 275
Romanian 50, 52–3, 98
Rongga 181–3, 185–8
Rotinese 176, 178, 181
Rotokas 109
Russenorsk 278, 299, 329
Russian 8, 13, 59–60, 94, 97, 99–101, 154–6, 169, 266–7, 269, 271, 275, 278, 290, 299

S
Samoan 292
Samoan Plantation Pidgin English 299, 329
Sango 275–9, 281, 299
Sanskrit 3, 210
Sanuma 275
Sãotomense 278, 298
Saparuan 180
Saramaccan 171, 173, 183, 278, 299
Semelai 274
Seychellois 278, 298
Shipibo-Konibo 71, 78, 80, 275
Sika 176–7, 179, 184, 187
Sirionó xi, 153, 155–61, 163–4
Sko 70, 78, 80
Slave 71, 75–8, 80, 85, 274
Solombala English 329
Solomons Pijin 299
Somali 70, 74–5, 79–80
Spanish 16, 49–50, 52–3, 156, 163, 174, 269–70, 274, 278, 280, 294, 299
Sranan 171, 278, 299
St. Lucian Creole 278, 298
Suena 275
Sundanese 120–2, 172–3

Supyire 274
Swahili 9, 19, 26, 146, 187, 275, 295
Swedish xiii, 161, 221, 244–5, 247, 254–6, 258, 275, 292, 321–2, 329, 331

T
Taba 275
Tagalog 275
Takia 232
Tauya 24
Tay Bôi 298
Tayo 278, 298
Tetun 175–7, 179–80
  Dili 175, 177, 179
  Terik 175–7, 179–80
Thai xi, 70, 75, 80, 110, 124, 133–7, 141, 144, 146, 148, 275
Timorese 177, 179, 181, 186
Tiwi 275
Tok Pisin 58, 179, 218–9, 222–3, 230, 233, 278, 299, 322
Tokodede 177–8
Trumai 71, 78, 80
Tukang Besi 178, 180, 184, 275
Turkish 9, 13, 269, 275, 294
Turku 298
Twi 120–1
Tzutujil 71, 80

U
Ukrainian 101
Umbundu 310–1
Urubú-Kaapor 71, 74–5, 78, 80, 159, 274
Usan 275

V
Vietnamese 9, 109, 120–1, 124, 129, 145, 270, 275

W
Warao 71, 79–80, 82, 84, 275
Wardaman 275
Wari' 274
Warlpiri 9
Warndarang 70, 79–80
Warunese 180
Waskia 231–2
Wavulu 288
Welsh 70, 80, 294
Wichí 275
Wichita 275

X
!Xóõ 109

Y
Yagua 71, 74–5, 78, 80, 275
Yamba 294
Yao 184–5
Yaqui 274
Yelî Dnye 70, 79–80, 84
Yidiny 275
Yimas 70, 79–80, 275
Yoruba 120, 121, 184, 185, 274, 310
Yukaghir, Kolyma, 274
Yurok 274

Z
Zoque, Copainalá, 274
Zulu 170, 274, 278
Zuni 274

Index of authors

A
Abbott, M. 156
Aboh, E.O. 213
Abraham, W. xiv
Aikhenvald, A.Y. 325
Aitchison, J. 10–1, 265
Allan, R. 247
Altmann, G. 47
Andersen, H. 4
Anderson, R.C. 321
Ansaldo, U. 213
Anttila, R. 33
Appleyard, D.L. 35
Arka, W. xiii, 167, 182, 185–6
Armbruster, C.H. 70
Arndt, P.P. 182–3
Augst, G. 247
Austin, P. 70

B
Babakaev, V.D. 299
Bailey, B. 277, 307
Bailey, T.M. 95
Baird, L. 181–2
Baker, M. 96
Baker, P. 277, 298
Bakker, P. xiii, 287, 297, 317, 321–3, 330
Baptista, M. 311, 315–6
Barrena, N. 277
Bartens, A. xiii, 305–6
Baxter, A.N. 277, 298, 310–1, 317
Becker, A.L. 125
Benedict, P.K. 134
Bergen, B. 269
Berger, H. 290
Bett, S. 49
Bhattacharjya, D. 299
Bickerton, D. 276, 307
Bisang, W. xiii, 135–7
Boas, F. 322–3, 325
Bollée, A. 277, 298
Booij, G.E. 248
Braun, M. 299
Braunmüller, K. 4
Bright, W. 35–6
Brito, A. de Paula 311–2, 315
Broadbent, S.M. 71
Broch, I. 277, 299
Brockman, J. 187
Brolin, M. 157
Busnel, R.-G. 268
Bybee, J.L. 195, 248, 251
Byington, C. 299

C
Campbell, G.L. 49, 53
Cardoso, E.A. 307
Carlson, L. xiii
Carrington, L.D. 277, 298
Carstairs-McCarthy, A. 139
Casad, E.H. 71
Castro, I. 309
Celce-Murcia, M. 140
Changizi, M.A. 44–5, 54
Chaudenson, R. 277, 282, 307
Cheng, X. 133
Chesterman, A. 141
Childs, G.T. 70
Chipere, N. vii
Chiravate, B. 146
Chomsky, N. 9, 28, 280
Christophory, J. 247
Clark, E.V. 260
Classe, A. 268
Cnattingius, A. 157
Cole, P. 71
Coleman, L. 335
Comrie, B. 68, 72, 290
Conklin, N.F. 136–7
Conrad, R.J. 70, 77
Corne, C. 277
Corsetti, R. 269
Court, C. 70
Covitt, R. 140
Cramer, I.M. 56
Croft, W. 31, 77, 83
Crowley, T. 295
Crystal, D. 29, 67, 265
Cysouw, M. xiii

D
Dahl, Ö. vii–viii, xi–xii, 4–5, 24, 27, 37, 67, 71–2, 85, 133, 135, 141–3, 145, 153–4, 162, 168, 191–3, 196, 202, 207, 214, 221–2, 245, 253, 321–2, 336–7
Dahlstrom, A. 70
Dammel, A. xii, 243–4, 317–8
Damoiseau, R. 277
Dayley, J. 71
De Groot, C. xii, 191–2, 195, 198, 202–4, 206–7, 212–4, 317
de Rooij, J. 247
de Vries, L.J. 70, 288
DeCamp, D. 307
DeGraff, M. 134, 268, 276–7, 282
Derbyshire, D.C. 71, 156
Derrmann-Loutsch, L. 247
Deutscher, G. vii, xiii, 23
Diab, M. 94
Diessel, H. 273
Dijkmans, J.J.M. 293
Dik, S.C. 212
Dixon, R.B. 293
Dixon, R.M.W. 28, 37, 156, 268
Dol, P. 70
Donaldson, B. 247
Donohue, M. 70, 167, 178–80
Drabbe, P. 178
Drechsel, E.J. 299
Dressler, W. 248
Dryer, M.S. 24, 69, 295
Du Bois, J. 73
Duke, J. 243
Dul'zon, A.P. 296
Durie, M. 173

E
Efron, B. 11
Ehrhart, S. 277, 298
Eisenberg, P. 247
Enfield, N.J. 111, 124, 331
Epstein, R. 141
Erickson, L. 217–8, 222
Eriksson, G. 217
Errington, J. 110
Evans, V. 57
Everett, D.L. VII, 71

F
Faraclas, N. 299
Fenk, A. X, 11, 24, 43, 46–7, 53, 55, 63, 73, 85, 110
Fenk-Oczlon, G. X, 11, 24, 43, 46–7, 53, 55, 63, 73, 85, 110
Fenyvesi, A. 191–2, 210
Ferraz, L.I. 277
Foley, W.A. 70, 135, 167, 178, 195
Fortescue, M. 70, 78
Frajzyngier, Z. 34
Frank, P. 71
Fromkin, V. 265
Fyle, C.N. 299

G
Gadelii, K.E. 5
Geerts, G. 247
Gell-Mann, M. VII, 25, 44–5, 62
Gibbs, G. 325, 327, 330–1
Gil, D. XI, 56, 109, 113–5, 124, 128, 134, 142–3, 145, 167, 171–5, 324, 331, 337
Giraud, G. 299
Givón, T. 68, 162
Gladwell, M. 148
Gordon, L. 70, 79
Gordon, R.G. Jr. 134, 220
Goyette, S. 187
Grant, A. 322–3, 325, 327–8, 330
Greenbaum, S. 247
Greenberg, J. 280
Grimes, C.E. 176
Grinevald, C. 182
Guirardello, R. 71
Günther, W. 277, 298
Guy, J.B.M. 299

H
Haan, F. de 325
Haan, J.W. 70
Haas, M. 144, 146
Haeseryn, W. 247
Hafford, J.A. 288
Hagman, R.S. 70
Hajek, J. 175, 179
Hale, H. 323
Hall, R.A. 58, 298–9
Hammarström, H. XII–XIII 287–8, 321
Hansen, Z. 247
Hansson, I.-L. 183
Hardman, M.J. 71
Harley, H. 273
Harris, A. 70
Harris, B.P. 322–3
Harris, J.B. 299
Haspelmath, M. VII, XII, 28, 72, 77, 269
Hawkins, J.A. VII, IX, 25, 37, 67, 72, 141, 245, 322, 337
Heath, J. 35, 70
Heimbach, E.E. 136
Heine, B. XIII, 58, 191, 209–11, 214, 277, 290
Henderson, J. 70
Hengeveld, K. 212
Hewitt, G. 266
Hinchcliffe, I. 247
Hock, H.H. 210
Hockett, C. 29, 265
Holm, J. 309, 317, 322–3
Holmes, P. 247
Holton, J. 139, 299, 323, 328–9, 332
Honório do Couto, H. 298
Hopper, P.J. 139
Hovdhaugen, E. 292
Huang, C.-T.J. 124
Huber, M. 317
Hull, G. 167, 177–8, 180–1
Hurd, C. 34
Hurd, P. 34
Huttar, G.L. 299
Huttar, M.L. 299
Hyams, N. 5
Hyman, L.M. 134, 184–5

I
Ingkaphirom, P. 70, 75
Iwasaki, S. 70, 75

J
Jacobs, M. 323, 325, 328
Jacobsen, J. 247
Jagacinski, N. 141
Jahr, E.H. 277, 299
Jake, J.L. 305, 307–8
Jarkey, N. 142
Jaya, I. 171, 178
Jespersen, O. 59
Johns, B. 144
Johnson, S.V. 323, 330–1
Jones, E.D. 299
Jonker, J.C.G. 176
Jordens, P. 260
Joseph, B.D. 210
Jourdan, C. 299
Juola, P. X, 25, 28, 44, 89, 93–6, 99
Jusayú, M.A. 294
Juvonen, P. XIII, 11, 321

K
Kakumasu, J. 71, 74, 159
Kaltenbrunner, S. 299
Karlsson, F. VII, 23, 67
Keesing, R. 276
Keller, R. 247
Kephart, R.F. 299
Kibrik, A.E. 109
Kihm, A. 182, 277
Kimball, G.D. 71, 75
King, G. 70, 101, 157, 183, 294, 326
Kiparsky, P. 111
Kittilä, S. 67, 76
Klamer, M. 176
Klein, W. 4
Klingler, T.A. 298
Koehn, E. 156
Koehn, S. 156
Kondangba, Y. 299
Köpcke, K.-M. 254
Koptjevskaja-Tamm, M. 23, 321
Kouwenberg, S. 71, 277
Kovács, M. XIII
Krashen, S. 145
Kress, B. 247
Kroch, A.S. 111
Kroeber, A.L. 293
Kulonen, U.-M. XIII
Kürschner, S. XII, 243–4, 317–8

Kusters, W. vii, ix, xii, 3, 10, 13, 16, 18–9, 23, 25–9, 32–3, 36, 38, 67, 84–5, 140, 145, 217, 221, 245–6, 248, 260, 267–8, 306, 318
Kutsch Lojenga, C. 70
Kuteva, T. 191, 209–11, 214

L
Labov, W. 307
Ladefoged, P. 45, 49
Lakämper, R. 6–7
Lang, J. 311
LaPolla, R.J. xiii, 70, 76, 137
Larsen-Freeman, D. 140
Laycock, D.C. 288, 299
Lean, G.A. 288
Lee, R. 217–8, 222
Lee-Smith, M.W. 179
Leech, G. 247
Lefebvre, C. 213, 276, 331
Lehmann, C. 111
Lehmann, W. 54
Lempel, A. x, 92–3, 106–7
Leplae, P. xiv
Lewis, E.D. 176
Lewis, G.L. 294
Li, C. 136, 138, 142–4, 147
Li, M. 24, 91, 290
Lindström, E. xii, 27, 70, 145, 217–8, 229
Liskin-Gasparro, J.E. 145
Lopes, N. de Silva 309–11, 316
Lorenzino, G.A. 298
Lucia, N.N. 294
Luffin, X. 298
Lundskær-Nielsen, T. 247
Lyons, C. 141
Lyovin, A.V. 136

M
MacDonald, L. 24
Maddieson, I. 12, 49, 55, 68, 109
Markey, T.L. 247
Maslova, E. 37
Massey, J.L. 92
Matisoff, J.A. 134, 136, 143–4
Mayerthaler, W. 248
McGregor, W.B. 70, 289
McSwain, R. 232
McWhorter, J.H. vii, xi, 4, 23–4, 27, 29–33, 36, 55, 67, 84–5, 89, 93–4, 104, 106–7, 134, 136, 145, 147, 167–9, 172, 181–3, 245, 265, 276, 283, 297, 305, 317, 322
Meintel, D. 307
Mello, H.R. de 316
Menzerath, P. 43, 46–8, 50–1, 55–6
Meyer, P. 47, 56
Miestamo, M. x, xii, 4–5, 23–4, 30–1, 33–4, 37–8, 55, 67, 69, 71–2, 221, 245, 259, 273
Mithun, M. 266, 323–4, 328, 331
Mosel, U. 292, 299
Moussay, G. 173
Mufwene, S. 134, 170, 277
Mühlhäusler, P. xiii, 4, 57–8, 268, 299, 321–2, 327, 329–31, 337
Mulaik, S.A. 50
Müller-Gotama, F. 173
Munteanu, D. 277, 299
Murane, E. 70, 79
Murray, E. 277
Muysken, P. xiii
Myers-Scotton, C. 305, 307–8, 311, 317

N
Næss, Å 68, 77
Nagaraja, K.S. 70
Nagel, T. 9
Nagy, W.E. 321
Nakayama, T. 71
Neef, M. 247, 252, 255
Nerbonne, J. 106
Neubauer, P. 133
Neumann, I. 298
Newmeyer, F.J. 37–8
Nichols, J. vii, xii–xiii, 74, 89, 100, 213, 251, 273
Niemi, J. xiii
Noonan, M. xiv
Nordlinger, R. 175
Norman, J. 266, 269
Nübling, D. 245, 247, 249

O
Olbertz, H. 195
Oldendorp, C.G.A. 299
Olsen, M.B. 94
Olson, M. 135
Omaggio, A.C. 145

P
Pagliuca, W. 195
Panagl, O. 248
Parker, G.W. 295
Parkvall, M. xii, 31, 134, 265, 287, 299–300, 321–2, 329
Pasco, D. 324–5, 327, 330, 335
Payne, D.L. 71, 74
Payne, T.E. 71, 74
Penchoen, T. 70
Perdue, C. 4
Perkins, M.R. 62
Perkins, R.D. vii, 195
Petersen, H.P. 247
Pinto, M.A. 269
Polikarpov, A.A. 60
Posner, R. 268
Prentice, D.J. 175
Press, I. 294
Priest, P. 157
Prokosch, E. 298
Pullum, G.K. 156

Q
Quint, N. 311, 316, 318
Quintero, C. 71
Quirk, R. 247

R
Radnóti, M. 209
Raidt, E.H. 247
Ramsey, S.R. 134
Ranta, A. 287
Ratliff, M. 134, 137, 140
Raukko, J. 56–7
Refsing, K. 293
Reinecke, J.E. 298
Rescher, N. viii–ix
Resnik, P. 94
Reyes, K.C. xiv
Rice, K. 71, 76
Richardson, I. 293
Rickford, J.R. 307
Riddle, E.M. xi, 11, 110, 133, 135, 137–9, 142
Riiho, T. xiii
Rijkhoff, J. 203, 205
Ritter, E. 273
Rodman, R. 265

Romero-Figueroa, A. 71
Romijn, K. 247
Ross, A. 299
Ross, M. 220, 229, 232
Rounds, C. 70
Rungruang, A. 133

S
Saeed, J.I. 70, 74
Saint-Fort, H. 282
Sampson, G. xiii
Sandström, N. xiii, 305
Sapir, E. 3, 134, 268
Scantamburlo, L. 277
Schaub, W. 70, 82
Scherre, M.M. Pereira 309
Schneier, B. 92
Schuchardt, H. 298
Schwegler, A. 174
Seifart, F. 182
Seiro, T. xiv
Shackle, C. 291
Shannon, C.E. 24, 62, 90–2, 95, 106
Shaw, G.C. 327, 330–1
Shnukal, A. 299
Shosted, R.K 23, 84–5, 110
Siewierska, A. 111
Silverstein, M. 322–3, 328, 336–7
Simon, H.A. 43, 45, 56, 160
Sinnemäki, K. x, 12, 23, 32, 67, 110
Smedberg, A. 321
Smith, N. 265
Smith-Hefner, N.J. 172
Sneddon, J.N. 70
Söderberg, B. 294
Spriggs, M. 217
Sridhar, S.N. 70, 74, 77
Stahlke, H. 135
Stassen, L. 31, 156, 158–9
Stefansson, V. 299
Strecker, D. 143–4
Sundblad, M. xiv
Svartvik, J. 247

T
Taylor, B. 116
Terrill, A. 70
Thao, L. 133
Thao, P. 133
Thepkanjana, K. 142
Thiele, P. 307
Thisted, R. 11
Thomason, S.G. 322–3, 328
Thompson, S.A. 136, 138–9, 142–4
Thráinsson, H. 247
Thurston, W.R. 4, 220–1, 231
Tiersma, P.M. 247
Tilman, A.V. 179
Tobler, A.W. 298
Todd, L. 57–8
Tolomeo, M. 269
Triwigati, Y. 116
Trudgill, P. 4

V
Vaes, K. xiv
Valenzuela, P.M. 71
Van de Kerke, S.C. 6–7
van den Toorn, M.C. 247
van der Voort, H. 24
Van Klinken, C.L. 175
van Name, A. 299
Van Valin, R. 195
Vang, L. 133, 143
Veiga, M. 298, 306–7, 311
Vendler, Z. 141
Vennemann, T. 82
Verhaar J.W.M. 277, 299
Versteegh, K. 269
Veselinova, L. 273
Vitányi, P. 24, 91, 290
Volker, C.A. 218, 222, 229
Vries, L. de 70, 288

W
Walsh, M.J. 295–6
Watkins, L.J. 34
Wegener, H. 260
Werner, S. 83
Widman, R. 294
Williams-Van Klinken, C. 175
Wilson, W.A.A. 298
Winford, D. 4, 317
Wittgenstein, L. 56
Wodehouse, P.G. 331, 335
Wogiga, K. 70, 77
Wray, A. 62
Wullschlägel, H.R. 299
Wunderlich, D. 6–7
Wurm, S.A. 179, 299
Wurzel, W.U. 243, 248, 318

X
Xiong, L. 133

Z
Zamponi, R. 295
Zenk, H. 323, 325
Zipf, G.K. 57, 90–1, 106
Ziv, J. x, 92–3, 106–7
Zubiri, J.O. 294

Index of subjects

Note: when there is no comma between a subentry and the following page number(s), the subentry should be read as main entry plus subentry, e.g., the combination of the entry "affix" and the subentry "order" reads as "affix order". Commas between subentries and page numbers indicate that this reading does not apply, e.g., the combination of the entry "agreement" and the subentry "gender," reads as "gender agreement", and the combination of the entry "animacy" and the subentry "human/animate distinction," indicates that the subentry belongs to the domain of the main entry.

4-M xiii, 305, 307–8, 310–3, 316–8

A
ablative 100, 110–2
accusative 74, 100, 110, 140, 176
adjective 59, 140, 202, 206, 208, 225, 231, 242, 280, 282, 309, 324, 329, 331–2, 334, 337
adverb 59, 141, 168, 180, 224, 237, 323–4, 332–4
affectedness 68
affix 7, 13, 17–9, 32, 35, 68, 133, 139–40, 147, 172–5, 178, 183–5, 210, 227–8, 230, 270
  see also loss of affix(ation)
  order 13–4, 32, 237
affixation 77, 109, 133, 135, 143, 145, 148, 172, 178–9, 181–2, 184–185, 226–7, 230, 233, 242, 260
agent 68, 76, 78–84, 98, 115, 118, 120, 171, 198–9, 201
agentive 180, 206
agglutinative 19, 35, 179, 185
  exponence 248
  language 116, 146, 179, 184–5, 210, 331
  morphology 48
  typology 184
agreement 32, 74, 81, 94, 114, 116, 199, 213, 226, 251, 253, 267, 272, 309
  see also agreement hierarchy, cross-reference
  gender, 136
  number, 309, 311
  object, 13, 17, 19, 83, 210, 225
  person, 68
  plural, 158, 204, 306
  redundant, 14, 25
  subject, 17, 83
agreement hierarchy 83
algorithm 91–2, 290
  see also compression algorithm
algorithmic information content (AIC) 44
alienability 175, 182–3, 270, 272
allomorph(y) xii, 7, 13–4, 32–3, 221, 243–8, 250, 252, 256, 258–60, 311, 317–8
allomorph assignment 243, 247, 252, 259, 318
'also' strategy 153, 159–60
ambiguity 60–1, 72, 84, 111, 113, 115, 146, 221, 251, 295
analytic(ity) xii, 170, 173, 177–9, 181–6, 267, 282, 300
  see also isolating
  expression 191–2, 194, 196–7, 200, 213–4
  form 194–6, 202
  language 49, 133–6, 143–7, 167, 169–70, 172–3, 178–9, 181, 183, 187, 323
  modality 209
  typology 180
animacy 72, 77–9, 119, 136, 193, 253–5, 311, 315–8
  human/animate distinction, 136, 311, 315–8
  inanimate, 72, 77–9, 119, 136, 254
antipassive 270, 272
applicative 272
argument x, 67–8, 72–9, 82–5, 119, 122–3, 173, 178, 198–201, 225, 227, 329
  marking 67–8, 72–7, 81–5
article 140–1, 226–8
  definite, 118, 153, 175, 270–1, 308–9, 330, 334
  indefinite, 271, 334
aspect(ual) viii, 14, 27, 31–2, 34–5, 68, 79, 115, 126, 128, 135, 140, 142, 146–8, 171–2, 186, 195, 224, 228, 280, 289
assignment 221, 226, 248, 253–9, 271, 318
  see also allomorph assignment, lexicalized assignment, semantic assignment
  principle 244, 246, 253–5, 259, 318
assimilation 251–3, 255–6, 259, 267
association experiment 116–7, 119–24, 129
association rule 115–6

associational 56, 109, 112–7, 119–25, 128–9, 142, 172–3, 324
  interpretation 56, 117, 119–24
  meaning 142
  semantics 56, 109, 112–6, 122, 172
asymmetry between affirmation and negation 37, 273

B

back-formation 199
baseline sociolinguistic profile 120–1
Bible x, 94, 97, 101, 106, 157–8, 288
bilingual(ism) 179, 219, 288, 307, 325
body-tally counting 288
bound
  case markers 140
  constructions 178
  declensional affixes 140
  forms 186, 270
  inflections 140, 147
  morpheme 135, 186, 196
  morphology xi, xiii, 321, 329–30
boundness 182, 185, 280

C

case 46–8, 68, 72, 78, 98, 100, 110–2, 139, 153, 172, 206, 208, 247–8, 251–2, 257–8, 267, 270, 272, 280, 282
  see also loss of case
  exponent 48
  marking 72, 74, 111, 114, 116, 123, 140, 148, 248, 258, 267
  system 89, 140, 258
categorical flexibility 322
causative 175, 184–5, 194, 200–1, 209, 214, 236, 325, 328
choice structure 37, 193–4, 197, 222, 337
Chomsky hierarchy ix
classifier xi, 78, 133–40, 144, 148, 171, 173, 175–6, 180, 182, 267, 270, 272, 296
clausal modifier 323, *see also* modifier
clitic 73–4, 111, 173–5
  see also enclitic, proclitic
coding-decoding asymmetry 250
coding domain 73, 82
coding strategy x, 67–8, 72–7, 81–2, 84
cognitive activation xiii, 307
cognitive cost 57, 59
collocation 61, 93, 119, 123, 183, 334
comitative 115, 158–9, 272
compensation viii, 29, 31, 55, 58, 68, 109–12, 124–6, 129, 134–5, 148, 251, 267, 269–71, 273
  see also complexity (counter)balance, complexity trade-off
  hypothesis xi, 109–12, 116, 122–3, 129
complexity
  see also elaboration, simplicity, word order complexity
  absolute, ix–x, xii, 3–5, 8, 13, 20, 23–9, 32–9, 72, 153, 211, 221, 244–6, 250–1
  (counter)balance 10, 47, 54–5, 57–8, 61–3, 67–8, 89, 110–1, 135, 142, 167, 267, 271
    see also compensation, complexity trade-off
  change in, xii, 12, 19, 210
  compositional, viii
  computational, viii
  conceptual, 133, 193, 205
  constitutional, viii–ix
  descriptive, viii–ix
  effective, 25, 44
  equal, viii, xi, 10–2, 23, 29, 31, 54–5, 63, 67–8, 85, 89, 101, 104, 106, 109–10, 130, 133, 256, 265–9, 282, 294
  equi-complexity-hypothesis, 31
  formal, 112, 248, 252
  formulaic, viii
  functional, ix
  generative, viii
  global, 23, 29–31, 39, *see also* complexity, overall
  hierarchical, ix
Kolmogorov, 24–5, 44, 91–2, 105, 290
linear, 92, 94–5, 105–7
linguistic, ix, 4, 45–6, 89–90, 93–4, 96, 101, 106, 130, 133, 145, 153, 162, 191–4, 213–4, 217, 245, 305–6
local, 29, 31, 39
LZ (Ziv-Lempel), 92–5, 105
measure(ment) ix, 8, 14, 24, 26, 28–9, 32, 38, 44, 89, 91–5, 99, 105–7, 112, 116, 133–4, 154, 193, 244–5, 270, 282
metric x, 30, 67–8, 72, 77, 92–4, 140, 145, 243–7, 259–60
morphological, 28–30, 58–61, 63, 98–9, 104, 111–2, 191–2, 210–4, 243–6, 256, 258, 260
nomic, ix
operational, ix
organizational, ix
overall, 4, 23, 29–31, 39, 54–5, 58, 63, 93, 106, 109, 112, 124–5, 129–30, 148, 163, 173, 183, 256, 259, 270, 296
pragmatic, 101, 104, 107, 125
psycholinguistic, 183
relative, ix–x, xii, 3–5, 8–14, 19–20, 23–8, 32–3, 37–9, 145, 221, 244–6, 250, 259, 290, 305
semantic, xi, 43, 55–60, 62–3, 109, 112, 116, 122, 124, 129
structural, 153–4, 163, 193, 205, 213, 220, 265, 281, 283, 336–7
syllable, 43, 46–9, 51–5, 68, 85, 271
syntactic, vii, 30, 89, 104, 106, 111–2, 124, 129, 163, 213, 317
system, 153–4, 163, 193, 196, 199–200, 202, 207, 213, 221, 245, 336
taxonomic, viii–ix
trade-off x, 43, 50, 54–5, 61, 63, 67–8, 84
  see also compensation, complexity (counter) balance

Compositional Matrix Language 308
compositional mode of interpretation 162–3
compositional rule 337
compositional semantics xi, 109, 112–6, 122–4, 129
compositionally associational/articulated 114–6, 119–25
compound(ing) xi, 118, 133, 135, 137, 143–4, 148, 194, 201–5, 209, 211, 327–30
compression algorithm 25, 28
condensation 153, 162
conjugation 6, 78, 94, 176, 169, 317
conjunction 156, 272, 334
construction 35, 46, 59, 104, 112–3, 115–22, 128, 141, 143, 147–8, 154–8, 161–3, 176, 178, 184–5, 191, 193, 195–6, 199–205, 207, 210, 212–4, 222, 224, 227, 229, 231, 266, 269, 280, 324, 328–9, 332–4, 337
constructional iconism 248
contact vii, ix, xi–xii, 3, 14, 19, 25, 31, 147, 168–9, 172, 174, 179–80, 184, 187, 191–2, 203, 209–10, 213–4, 219, 230, 277, 281, 310, 317, 321–2, 325
  variety xii, 170, 174, 317
contact-induced simplification 27, 29, 31
content morpheme 307–8, 311–3, 317
context(ual) 30, 37, 44, 56–7, 59, 62, 68, 72, 77, 82, 84, 91–2, 96, 98, 101, 109, 113–5, 124–6, 128, 139–41, 147–8, 167, 174–5, 186, 194, 196, 204, 227, 248–9, 253, 294, 311, 315, 323–4, 328, 337
co-occurrence restrictions 148, 229
coordination 141, 153, 155–8, 161, 163
  see also 'also' strategy, list-strategy, 'with' strategy
copula 171, 175, 208, 212–4, 270, 272, 332
core argument, *see* argument
core argument marking, *see* argument marking
cost (and difficulty) ix–x, 23–9, 37–9, 268
cost-benefit trade-off 267
creole vii, xi–xiii, 20, 29, 31, 55, 69, 71, 104, 107, 120–1, 134, 136, 145, 167–9, 171, 173–4, 179, 182–4, 218, 222, 265, 276–9, 281–3, 287, 290, 297–300, 305–8, 311, 317, 325, 330–1, 336
  first and second generation creole 307
  formation 305, 318
creolization xiii, 168, 174, 305–8
cross-linguistic comparability x, 23, 31, 258–60
cross-linguistic correlation 31, 43, 45–7, 50–1, 54, 59, 63
cross-linguistic dispensability 143
cross-linguistic frequency 37–8, 83, 269–70
cross-linguistic rarity x, 23–4, 37–9
cross-reference 35, 223–5, 228, 233
cultural context 96

D

definiteness 115, 126, 128, 175, 221, 308, 311, 315–8
  definite 72, 77–8, 141, 317
  indefinite 72, 77, 141
demonstrative 137, 148, 226, 270–2, 323
dependent marking x, 67–8, 73–84, 251, 280
derivation 143, 194, 206–7
  syntactic, *see* syntactic derivation
derivational viii, 168, 181, 198, 206, 253, 329
  history 193, 199, 205, 207
  morphology xiii, 10, 173, 198, 321, 327, 329–30
  rule 198, 206, 214, 221, 329, 337
description language 289–90, 294–5
description length ix, xiii, 25, 27–8, 32, 34, 36, 38, 250–1, 290, 296
determiner 140, 251, 309–10, 312–4
dialogue 325, 335
difficult(y) vii, x, xii, 8–9, 12–4, 23–9, 33, 37–9, 72, 101, 140–1, 145, 147, 153, 177, 217, 219–22, 224–33, 245–6, 260, 265
direction of determination between stem and suffix
  left-to-right 255–6, 259
  right-to-left 255, 259
distinctiveness x, 67, 71–2, 78, 84–5, 148
double-object 280
dual 126–7, 158, 224–6, 241, 270

E

economical numeral expressions 296
economy x, 13, 32–3, 36, 55, 63, 67, 71–2, 84–5, 90, 135, 295–6
efficiency ix, 37, 267–8, 321, 336–7
  see also difficulty, inefficiency
elaborate expression xi, 143–5
elaboration 109–10, 115, 124, 142, 144–5, 148, 178, 270
  see also grammatical elaboration, lexical elaboration, structural elaboration
enclitic 175, 177, 224, 242
ergativity 173, 178–9, 183, 323
Ethnologue 134, 220, 280
etymology 161, 291, 297, 322, 330
Eurocentric 3–4, 126, 268, 276, 331
evasion 228
evidential(ity) 30, 32, 36–7, 167, 183, 270–2, 325
exceed-comparatives 280, 282
experiment(al) xi–xii, 33, 46, 93, 95, 98, 101, 104, 106, 109, 134, 141, 246
  see also association experiment
expressivity 44, 245, 269

extra-linguistic 124–5, 128, 148, 232–3

F
feminine 136, 225, 227, 250, 254
focus 68, 78–9, 178, 197
form-meaning relationship 34–7, 146, 243–4, 248, 252, 260
form transparency 229–231
free morpheme 135, 143, 175, 178, 183, 225
free variation 71, 196
frequency 57, 59, 72, 77, 250–1, 289, 337
  see also cross-linguistic frequency
  asymmetries 72
functional domain x, 23–4, 31–4, 36–7, 39, 68, 72–3
functional level 308
functional load x, 67, 71–3, 76–82, 84, 289
fused encoding 247–8, 252, 260
fusional 48–9, 185
future tense 34, 78, 195, 224, 227, 229, 272, 323

G
gender 35, 46, 68, 75, 78, 89, 94, 139–41, 148, 153–4, 168, 182, 221, 223, 225, 228, 230, 232, 252–5, 259, 267, 271–2, 318, 323–4
  see also agreement, feminine, masculine
generative grammar 5, 28, 37–8
genus 69–71
grammatical elaboration 145, 148
grammatical resources xi, 153–5, 163, 245, 336–7
grammatical rule 115, 207, 323
grammaticalization xi, 31–2, 34, 36, 38, 111, 142, 146, 153–4, 158, 161–2, 167–8, 170, 174, 178, 185, 253, 271–2, 334

H
Halting Problem 92
head 75, 77, 115, 156, 172, 178, 182, 207, 212, 251, 308, 312, 323, 332
  marking x, 67–8, 73–84, 251, 280
hearer 12, 14, 25–6, 72, 78, 90–1, 96, 105, 128
heterosemy 331
hierarchy 43–5, 192, 212–3, 282
  see also agreement hierarchy, Chomsky hierarchy, hierarchical complexity, person hierarchy
homonymy 13–4, 26, 32, 43, 46, 56–61
homophony 36, 57, 59, 72, 78, 140, 144, 161

I
iconic(ity) 113, 115, 243, 246, 248, 250, 252, 317–8
idiolect(al) 57, 325
idiomatic speech 43, 58–61
inanimate, *see* animacy
incorporation 323
incremental (mode of interpretation) xi, 162–3
independent morpheme 143
inefficiency 38, 72
infix(ation) 173, 242
inflection x, 6, 10, 13, 17–20, 25–6, 32–3, 89, 93–5, 133–5, 140–1, 146–7, 170–1, 178–80, 182, 185, 194, 323
  verbal, *see* verbal inflection
inflectional 6, 10, 14, 17, 178–9, 185, 210, 248, 253
  affix(ation) 32, 133, 140, 172, 185
  change 15–6, 20
  marking xiii, 305, 317
  morpheme viii, 143, 147
  morphology 3, 10, 12–5, 89, 94, 135, 147, 178, 206, 259, 306
  paradigm viii, 89, 95
  rule 214
  synthesis 85
information theory 24, 27, 44, 61–2, 72, 90, 96, 105–7, 134, 140, 336–7
interdependency 226
interrogative 137, 196, 323
intransitive 154, 198–9, 209, 222, 242
irregularity 13, 29–30, 32–3, 36, 94, 106, 167, 176, 184, 247, 249, 268
  in paradigms 220–2, 226, 233
  of numerals 290–1, 294, 296
isolating 49, 170, 210
  see also analytic
  language xi, 56, 109–10, 112, 116, 118, 120–5, 128–30, 133–4, 176–9, 324, 331
  morphology 48, 60, 135
  vs. non-isolating 109, 112, 116, 120–4, 129–30
isomorphy 13–4, 32–3

J
juxtaposition 113–5, 119, 134–5, 140–3, 156, 159, 300, 324

L
language acquisition 5, 11, 16, 26, 62, 133, 140, 145, 213, 221, 232, 246, 260, 316–7
  by adults, 167, 221, 246
  by children, 5, 213
  first (L1) language, 26, 133, 246, 260
  foreign language, 11, 140, 145
  non-native, xi, 167–70, 174, 180, 186–7
  second (L2) language, 11, 16, 145, 175, 246, 260, 316–17
language acquisition device 5
language change vii–viii, xi, 3, 19–20, 54, 71–2, 85, 136, 148, 170, 174, 179–81, 185, 187, 191, 195, 204, 209–10, 213–4, 300, 317
language contact, *see* contact
language ecology 219
language learner 19, 26, 141, 146–7, 179, 217, 221–2, 225, 228–33, 246, 260
  adult learner, xii, 8, 25–6, 217, 221–2
  first language (L1) learner, 9, 13–4, 25–6
  foreign/second language learner, (L2) x, 4, 9,

12–6, 19, 25–7, 33, 140, 147, 222, 245, 260
language learning viii, 13, 85, 145, 172, 232
see also language acquisition
foreign, 11, 140, 145
language shift 231
language use 5–6, 11, 13, 63, 72, 145, 148, 221, 232–3, 243, 246, 259
language variety x, xiii, 3, 6, 15–9, 134, 168–172, 175–9, 191–2, 194–5, 197, 201, 207–12, 218, 268, 277–8, 289, 294, 297, 305–8, 311, 317, 321–31, 336–7
learnability 147, 220
lemma 307–8
length of derivational history, *see* derivational history
length of description, *see* description length
lexeme viii, 28, 56, 93, 154, 210–1, 227, 230, 292, 309, 327
lexical
elaboration xi, 133–6, 140, 145, 147–8
item xiii, 98, 104, 135, 140–2, 144, 148, 154–5, 206, 230, 321–2, 327, 330–1, 335–7
morpheme 56, 112, 321, 327–31
rule 322
semantics 10, 113
lexicalization 143, 198–9, 249, 265, 312, 328
lexicalized assignment 254
lexicon x–xi, xiii, 7, 11, 25, 28, 55–8, 68, 76, 92–3, 95, 105, 107, 110, 135, 154, 221, 245, 258–9, 308, 321–3, 325, 330, 336–7
lexifier xiii, 277–8, 287, 297–300, 305, 308, 322–3, 329–30
linear order(ing) 111, 114, 116, 122, 133, 142, 148, 162, 310
linearity 141, 154, 193, 196–7, 213
see also non-linearity
lingua franca xi, xiii, 15–7, 19, 147, 170, 175, 218, 231, 278, 321
linguistic theory vii, 4–5, 8, 27, 33
list-strategy 153, 159–60
locative 119, 208, 270, 272, 323–4, 332
loss of affix(ation) 135, 147, 179
loss of case 111, 180
LZ-compression x, 94, 96, 105

M

macro-area 69
markedness vii, 5, 28, 72, 77, 248
theory 28, 248
masculine 136, 139, 225–8, 249–55, 271, 309
maturation viii, 134
meaning association 134
meaning compression 142
metatypy 232, 297
modality 194–7, 209, 214
modifier 115, 140, 182, 208, 212–3, 323, 332–4
monosyllable x, 43, 45, 48–54, 59–60, 63
morpheme 56, 112, 135, 143, 147, 175, 178, 183, 185–6, 196, 225, 307–13, 317, 321, 327–31
see also content morpheme, independent morpheme, system morpheme
inventory 154, 220–1
morphology viii, x, 3–4, 6–8, 12–3, 45–6, 60–1, 63, 67, 98–9, 109–12, 122–3, 127–30, 135, 147, 171–3, 178–9, 183–5, 219, 225–6, 248, 251, 256–60, 265, 271, 280, 323–4, 327, 329–30
morphophonological viii–ix, 140, 308
process 230
morphosyntax xiii, 34, 36, 75–6, 84, 113–6, 124–5, 127–9, 158, 196, 202, 214, 305, 308, 324
multifunctionality xiii, 321–2, 330–1, 334–7

N

narrative 325
Natural Morphology 248, 251, 318
negation 34–7, 173, 178, 180, 224, 229, 237, 272–3, 280, 323
nominal morphology 225–6
nominal phrase 244
nominative 74, 100, 110, 140, 176
non-linearity 154, 253
noun 126–7, 136–41, 153–4, 156–8, 161–2, 201–5, 208–9, 211–2, 222–3, 225–31, 243–4, 247–8, 250–1, 253–9, 270–1, 315–8, 323–4, 329–34, 337–8
class 78, 136, 140, 226, 270
classification 136, 139
phrase 73–6, 139–41, 156–63, 223–8, 250–1, 271–2, 309, 318, 329, 332–4, 338
number 115, 126, 139, 205, 224–42, 248–60, 309–17
agreement 309, 311
base 291–5, 298, 304
marking 126, 205, 221, 225, 243, 248–9, 256, 280
numeral 171–6, 180, 182, 204–5, 225, 267, 287–92, 294–300, 311–2, 329, 334
classifier 171, 173, 175–6, 180, 182, 267, 270, 272

O

object 7, 13, 17–9, 36, 60, 76, 78, 83, 111, 141, 171–2, 177–8, 180, 188, 198–9, 203–5, 209–11, 214, 223–5, 227–8, 233, 239, 269, 271, 280, 297, 314–7, 323, 332, 334
one-meaning–one-form viii, x, xii, 32–7, 39, 71–2, 85
orthography 135, 157, 173, 205, 229, 325–6, 330
output structure 193–4
overspecification ix, 29–30, 32–3, 36, 167, 173, 176, 182–3

P

paradigm 17–8, 35, 176, 220–4, 227, 231, 233, 317, 323
inflectional, *see* inflectional paradigm

paradigmatic 18, 32, 34–7, 147, 233
relationship 140
participant-oriented modality 195
particle 35, 133, 178, 183, 324
part-of-speech VIII, 59–61, 324, 328–9, 331, 334, 336
passive 68, 168, 171, 184, 186, 199, 266, 269, 272
past tense 17–8, 34, 37, 146, 323
patient 68, 77–9, 81, 84, 98, 115, 118–23, 171, 201
pattern regulation 196, 199
perfective 35, 78, 135, 146–7, 272
person 7, 13, 17–8, 35–6, 68, 74–9, 93, 168, 176, 186, 195–6, 200, 224–7, 237–8, 272, 280, 308, 323, 332
hierarchy 76
phenogrammatics 193, 211, 214
phoneme 12, 24, 45–54, 57–8, 268–71, 290, 296, 337
phonemic inventory 24, 46, 48–9, 53–5, 59, 61, 89, 109, 271
phonology 5, 10, 43, 45–6, 55, 58, 61, 63, 89, 127, 129, 185, 229
pidgin XII–XIII, 55–8, 63, 145, 180, 181, 218, 263–5, 277–8, 281–3, 287, 297–300, 321–37
lexicon 322, 327, 330–1, 336
pidginization 168, 174, 305, 321
plural XII–XIII, 7, 12–3, 17–8, 78–9, 98, 125–7, 137–40, 158, 174, 203–5, 209, 211, 214, 225–8, 243–9, 252–6, 260, 271, 305–6, 308–18
allomorphy 243, 245, 258, 311
polysemy 43, 46, 56–61, 230, 322, 327, 330–1
positional level 308
possession 15, 113, 137, 139, 171, 175–6, 178, 180, 182–4, 201, 203–5, 210, 213, 224–5, 227, 229, 231, 239, 241, 271–2, 311, 323–4, 332, 334
see also alienability
alienable, 184, 224–5, 227, 229, 241
inalienable, 183–4, 224–5, 229, 239
postposition 48, 74, 158, 180, 202, 208
pragmatic(s) XI, 10, 12, 30, 62, 68, 101, 109–10, 124–9, 145, 161–2, 171, 174, 183, 196–7, 271, 307
see also complexity, pragmatic
inference VII, 127
predicate 68, 74–6, 158, 162, 175, 200–1, 329, 337
prefix 36, 79, 162, 168, 173, 176, 180–1, 225–6, 228–30, 242, 280, 310
preposition 48, 59–60, 110–1, 224–5, 267, 308–9, 324, 332–3
present tense 34, 78–9, 323
principle of fewer distinctions 33–4, 36–7
principle of one-meaning–one-form, *see* one-meaning–one-form
principle of relevance 248
problem of comparability 23, 30–1, 259–60
problem of representativity 23, 30–1
processing VII, IX, 5, 9, 13–4, 25–7, 32, 36, 38–9, 72, 89, 95, 104–6, 221, 246
proclitic 75, 183
proficiency VII, 145–6
pronominal paradigm 180, 227
pronominal system 188, 224–6, 323
pronoun 72–3, 76, 78, 101, 124, 198, 224–7, 238–9, 271–2, 280, 289, 310–1, 323, 325, 332–4
system 226
pronunciation 141, 325
prosody 253–4, 329
psychological reality 28, 38

R

recursion 38, 127, 206, 304
redundancy 14, 25, 62, 71–2, 93–4, 112, 143, 248, 250–1, 257–8, 268, 271, 312, 316–8
reduplication 133–5, 144, 271, 327, 330, 337
referentiality 31, 68
reflexive 194, 197–9, 209, 213
regulations 153–5, 214, 245, 336–7
grammatical, 193
pattern, 196, 199
relational information 140
relational word 324
remapping 233
replica language 209–10
resultative construction 143
rule 61, 74, 114–6, 134, 198, 206–7, 214, 221, 322–3, 329, 337
number of rules 6, 30, 32–3, 38, 45, 62, 167, 269, 291
obligatory, 134
optional, 134

S

sample X, XII, 46–7, 49, 54–5, 67–70, 82–4, 92, 94, 98, 100–1, 104, 107, 111, 136–7, 192, 243, 245, 249, 276–9, 281, 298
scrampression 95
Seinsart 205, 211, 214
semantic(s) X, XII, 45–7, 56–7, 61–2, 94, 109, 125–7, 130, 134–5, 143, 156, 180, 186, 193, 231, 248, 254, 289, 311, 317, 324
see also associational semantics, compositional semantics, lexical semantics
assignment 253–4
indeterminacy 10, 114–5
interpretation 115–6, 324
organisation 322
rule 114–5
serialization 133, 135, 140–3
serial verb 141–2, 227
set (noun type) 203–5, 211, 317
simplicity VII–VIII, X, 10–1, 28–9, 31, 56–7, 67–8, 84, 112, 129–30, 146, 187–8, 258, 265, 267, 276, 279, 282, 297, 321–2, 324, 337
absolute 38–9

grammatical, 129, 167
morphological, 110, 112, 129
morphosyntactic, 125
of creoles vii, 265, 276, 297
overall, 109, 168, 173
perceptual, 11
word order, 63
simplification 14, 19, 25, 85, 255
singular 36, 74–5, 78–9, 93, 125–7, 140, 168, 176, 196, 204, 224–5, 227–30, 244, 249, 308, 311–7
object (noun type) 203–5, 209, 211, 214
situation-oriented modality 195
social status 145, 233, 288
speaker 8–11, 14–7, 19, 25–6, 37, 62, 72, 90–1, 96, 105, 145–6, 193, 221, 227–8, 246, 250, 307–8
speech community 9, 14–7, 287, 292, 297
stem involvement 247–50, 255, 258–60
structural asymmetry 72
structural elaboration ix, 29–30, 32–3, 36, 167
subject 7, 17, 36, 61, 78, 83, 101, 111, 154, 156, 158, 161, 172–3, 176, 178, 181, 196, 199, 224–5, 227–8, 230–1, 272, 280, 323, 332, 334
agreement 17, 83
substrate 57, 276, 307–8, 317, 322–3
subsystem 54–5, 58, 61, 63, 109, 147, 223, 226, 267
subtractive marking 251
suffix 35, 74–5, 98, 110–1, 118, 195–7, 200, 206, 224–6, 242–4, 247, 251–6, 259
suffixation 244–5, 247–50, 252, 310–1, 315
suppletion 79, 99, 183, 221, 233, 247, 249–50, 258, 272
syllable 6, 38, 45–55, 57–61, 63, 85, 135, 144, 184–5, 196, 251, 268–9, 271, 273, 296
see also monosyllable
complexity x, 43, 46–55, 68
structure 55, 57–8, 61, 85, 271, 273
type x, 43, 46, 48–52, 54, 59–61
syncretism 139, 225, 248
synonymy 71, 327
syntactic derivation 37–8
syntagmatic 34, 147
syntax viii, x, 7, 10, 28–9, 45–6, 55, 61, 63, 67, 98, 104, 109–12, 122–4, 127, 129–30, 139, 141, 162, 173, 197, 207, 226, 231, 233, 322, 324, 336
synthesis 85, 184, 270, 280, 282
synthetic 209, 270, 318
expression 191–2, 194–7, 200, 213–4
language 116, 133–5, 139, 143, 145–8, 282
system morpheme 307–10, 312, 317

T
tectogrammatics 193, 197, 211, 214
tense 7, 14, 17, 30–2, 34, 36–7, 39, 115, 126, 128, 148, 164, 171–2, 178, 186, 195, 200, 214, 223–4, 228, 237, 240, 280, 289
thematic role 111, 115, 120, 123–4, 126, 128, 308, 312
tone 55, 74, 79, 82, 129, 134–5, 148, 169, 267, 271
complexity of, 134
sandhi 140, 143–4
topic 73, 75, 82, 139, 146
continuity 73
trade-off x, 10–2, 31–2, 43, 50, 54–5, 61–3, 67–73, 77, 81–5, 124, 129, 267
hypothesis 68, 82–4
transitive 68, 74–5, 77, 178, 186, 197–9, 206, 209, 222–3, 227, 230, 242
transparency 7, 13, 32–3, 36, 229, 231, 248, 252
trial 126–7, 224–5, 228
Turing machine 105, 289
typology 3, 20, 23, 31, 51, 69, 77, 111, 114, 116, 129, 135–6, 143, 145, 148, 179–81, 185, 187, 191–2, 195, 202, 204, 209–14, 222, 244, 265–6, 269, 273, 276, 278–83
agglutinative, 184
analytic, 179–80,
associational, 114, 116, 122
fusional 185
overall, viii
systemic, x, 47

U
underdifferentiation 109, 126
underspecification 7, 113–5, 123
use pattern 191, 209–10, 214
uniformitarian 277, 279, 281
universal xiii, 5, 68, 96, 107, 111–2, 114, 192, 195, 307, 324
meaning representation system 125–8
semantics 125–7

V
vagueness 109, 113, 115, 127–8, 156, 237, 260, 290, 337
validation step 243, 246, 251, 259–6
verb 6, 13, 17, 59–61, 68, 75, 77–9, 85, 93–4, 111, 141–3, 147–8, 154, 168, 176, 184–6, 191–200, 204, 206, 215, 222–8, 230–1, 233, 236–7, 242, 280, 282, 317, 323–4, 328–9, 331, 334, 337–8
class 77–8, 198–9, 222, 224–8, 230–1, 237, 242
morphology 4, 17, 77, 271
phrase 156, 223–4, 226–8, 233
serialization 133, 135, 140–1
verbal inflection x, 3, 6, 10, 19–20, 25–6, 32–3
verbalizing morpheme 329
verbosity 134–5
vocabulary xiii, 57, 109, 154, 221, 229–31, 289, 298, 321–2, 325–30, 336

W
whistled languages 268
'with' strategy 153, 158, 160
word class 140, 248, 331
word formation 43, 50, 135, 143

word order ix–x, 8, 28, 38, 43, 46–7, 56, 60–3, 67–8, 73–5, 77–8, 81–2, 84, 94, 98, 111, 116, 120, 122–3, 135, 168, 171, 179, 212, 223, 227, 267, 269, 282, 323
see also affix order, linearity, linear order(ing)
canonical, 75, 82
complexity x, 61–3
flexible, 110–1, 122
OVS, 5
reversible, 75, 82
rigid 43, 46, 60–3, 111, 123, 135
rule 61, 74
SOV, 179, 212, 269, 280
SVO, 5–6, 8, 56, 116, 119–20, 223, 227, 231, 269, 280, 282, 323
VSO, 8, 223, 227, 231
World Atlas of Language Structures (WALS) xii, 69, 269–74, 276, 278–81

Z

zero marking 75, 213, 243, 245, 251, 254, 258
Ziv-Lempel compression, *see* LZ compression

Studies in Language Companion Series

A complete list of titles in this series can be found on the publishers' website, *www.benjamins.com*

100 **AMEKA, Felix K. and Mary Esther KROPP DAKUBU (eds.):** Aspect and Modality in Kwa Languages. vii, 344 pp. *Expected March 2008*

99 **HØEG MÜLLER, Henrik and Alex KLINGE (eds.):** Essays on Nominal Determination. From morphology to discourse management. xviii, 366 pp. + index. *Expected April 2008*

98 **FABRICIUS-HANSEN, Cathrine and Wiebke RAMM (eds.):** 'Subordination' versus 'Coordination' in Sentence and Text. A cross-linguistic perspective. vi, 351 pp. + index. *Expected April 2008*

97 **DOLLINGER, Stefan:** New-Dialect Formation in Canada. Evidence from the English modal auxiliaries. 2008. xxii, 355 pp.

96 **ROMEO, Nicoletta:** Aspect in Burmese. Meaning and function. 2008. xv, 289 pp.

95 **O'CONNOR, Loretta:** Motion, Transfer and Transformation. The grammar of change in Lowland Chontal. 2007. xiv, 251 pp.

94 **MIESTAMO, Matti, Kaius SINNEMÄKI and Fred KARLSSON (eds.):** Language Complexity. Typology, contact, change. 2008. xiv, 356 pp.

93 **SCHALLEY, Andrea C. and Drew KHLENTZOS (eds.):** Mental States. Volume 2: Language and cognitive structure. 2007. x, 362 pp.

92 **SCHALLEY, Andrea C. and Drew KHLENTZOS (eds.):** Mental States. Volume 1: Evolution, function, nature. 2007. xii, 304 pp.

91 **FILIPOVIĆ, Luna:** Talking about Motion. A crosslinguistic investigation of lexicalization patterns. 2007. x, 182 pp.

90 **MUYSKEN, Pieter (ed.):** From Linguistic Areas to Areal Linguistics. vii, 274 pp. + index. *Expected February 2008*

89 **STARK, Elisabeth, Elisabeth LEISS and Werner ABRAHAM (eds.):** Nominal Determination. Typology, context constraints, and historical emergence. 2007. viii, 370 pp.

88 **RAMAT, Paolo and Elisa ROMA (eds.):** Europe and the Mediterranean as Linguistic Areas. Convergencies from a historical and typological perspective. 2007. xxvi, 364 pp.

87 **VERHOEVEN, Elisabeth:** Experiential Constructions in Yucatec Maya. A typologically based analysis of a functional domain in a Mayan language. 2007. xiv, 380 pp.

86 **SCHWARZ-FRIESEL, Monika, Manfred CONSTEN and Mareile KNEES (eds.):** Anaphors in Text. Cognitive, formal and applied approaches to anaphoric reference. 2007. xvi, 282 pp.

85 **BUTLER, Christopher S., Raquel HIDALGO DOWNING and Julia LAVID (eds.):** Functional Perspectives on Grammar and Discourse. In honour of Angela Downing. 2007. xxx, 481 pp.

84 **WANNER, Leo (ed.):** Selected Lexical and Grammatical Issues in the Meaning–Text Theory. In honour of Igor Mel'čuk. 2007. xviii, 380 pp.

83 **HANNAY, Mike and Gerard J. STEEN (eds.):** Structural-Functional Studies in English Grammar. In honour of Lachlan Mackenzie. 2007. vi, 393 pp.

82 **ZIEGELER, Debra:** Interfaces with English Aspect. Diachronic and empirical studies. 2006. xvi, 325 pp.

81 **PEETERS, Bert (ed.):** Semantic Primes and Universal Grammar. Empirical evidence from the Romance languages. 2006. xvi, 374 pp.

80 **BIRNER, Betty J. and Gregory WARD (eds.):** Drawing the Boundaries of Meaning. Neo-Gricean studies in pragmatics and semantics in honor of Laurence R. Horn. 2006. xii, 350 pp.

79 **LAFFUT, An:** Three-Participant Constructions in English. A functional-cognitive approach to caused relations. 2006. ix, 268 pp.

78 **YAMAMOTO, Mutsumi:** Agency and Impersonality. Their Linguistic and Cultural Manifestations. 2006. x, 152 pp.

77 **KULIKOV, Leonid, Andrej MALCHUKOV and Peter de SWART (eds.):** Case, Valency and Transitivity. 2006. xx, 503 pp.

76 **NEVALAINEN, Terttu, Juhani KLEMOLA and Mikko LAITINEN (eds.):** Types of Variation. Diachronic, dialectal and typological interfaces. 2006. viii, 378 pp.

75 **HOLE, Daniel, André MEINUNGER and Werner ABRAHAM (eds.):** Datives and Other Cases. Between argument structure and event structure. 2006. viii, 385 pp.

74 **PIETRANDREA, Paola:** Epistemic Modality. Functional properties and the Italian system. 2005. xii, 232 pp.

73 **XIAO, Richard and Tony McENERY:** Aspect in Mandarin Chinese. A corpus-based study. 2004. x, 305 pp.

72 **FRAJZYNGIER, Zygmunt, Adam HODGES and David S. ROOD (eds.):** Linguistic Diversity and Language Theories. 2005. xii, 432 pp.

71 **DAHL, Östen:** The Growth and Maintenance of Linguistic Complexity. 2004. x, 336 pp.

70 **LEFEBVRE, Claire:** Issues in the Study of Pidgin and Creole Languages. 2004. xvi, 358 pp.

69 **TANAKA, Lidia:** Gender, Language and Culture. A study of Japanese television interview discourse. 2004. xvii, 233 pp.

68 **MODER, Carol Lynn and Aida MARTINOVIC-ZIC (eds.):** Discourse Across Languages and Cultures. 2004. vi, 366 pp.

67 **LURAGHI, Silvia:** On the Meaning of Prepositions and Cases. The expression of semantic roles in Ancient Greek. 2003. xii, 366 pp.

66 **NARIYAMA, Shigeko:** Ellipsis and Reference Tracking in Japanese. 2003. xvi, 400 pp.

65 **MATSUMOTO, Kazuko:** Intonation Units in Japanese Conversation. Syntactic, informational and functional structures. 2003. xviii, 215 pp.

64 **BUTLER, Christopher S.:** Structure and Function – A Guide to Three Major Structural-Functional Theories. Part 2: From clause to discourse and beyond. 2003. xiv, 579 pp.

63 **BUTLER, Christopher S.:** Structure and Function – A Guide to Three Major Structural-Functional Theories. Part 1: Approaches to the simplex clause. 2003. xx, 573 pp.

62 **FIELD, Fredric:** Linguistic Borrowing in Bilingual Contexts. With a foreword by Bernard Comrie. 2002. xviii, 255 pp.

61 **GODDARD, Cliff and Anna WIERZBICKA (eds.):** Meaning and Universal Grammar. Theory and empirical findings. Volume 2. 2002. xvi, 337 pp.

60 **GODDARD, Cliff and Anna WIERZBICKA (eds.):** Meaning and Universal Grammar. Theory and empirical findings. Volume 1. 2002. xvi, 337 pp.

59 **SHI, Yuzhi:** The Establishment of Modern Chinese Grammar. The formation of the resultative construction and its effects. 2002. xiv, 262 pp.

58 **MAYLOR, B. Roger:** Lexical Template Morphology. Change of state and the verbal prefixes in German. 2002. x, 273 pp.

57 **MEL'ČUK, Igor A.:** Communicative Organization in Natural Language. The semantic-communicative structure of sentences. 2001. xii, 393 pp.

56 **FAARLUND, Jan Terje (ed.):** Grammatical Relations in Change. 2001. viii, 326 pp.

55 **DAHL, Östen and Maria KOPTJEVSKAJA-TAMM (eds.):** Circum-Baltic Languages. Volume 2: Grammar and Typology. 2001. xx, 423 pp.

54 **DAHL, Östen and Maria KOPTJEVSKAJA-TAMM (eds.):** Circum-Baltic Languages. Volume 1: Past and Present. 2001. xx, 382 pp.

53 **FISCHER, Olga, Anette ROSENBACH and Dieter STEIN (eds.):** Pathways of Change. Grammaticalization in English. 2000. x, 391 pp.

52 **TORRES CACOULLOS, Rena:** Grammaticization, Synchronic Variation, and Language Contact. A study of Spanish progressive -ndo constructions. 2000. xvi, 255 pp.

51 **ZIEGELER, Debra:** Hypothetical Modality. Grammaticalisation in an L2 dialect. 2000. xx, 290 pp.

50 **ABRAHAM, Werner and Leonid KULIKOV (eds.):** Tense-Aspect, Transitivity and Causativity. Essays in honour of Vladimir Nedjalkov. 1999. xxxiv, 359 pp.

49 **BHAT, D.N.S.:** The Prominence of Tense, Aspect and Mood. 1999. xii, 198 pp.

48 **MANNEY, Linda Joyce:** Middle Voice in Modern Greek. Meaning and function of an inflectional category. 2000. xiii, 262 pp.

47 **BRINTON, Laurel J. and Minoji AKIMOTO (eds.):** Collocational and Idiomatic Aspects of Composite Predicates in the History of English. 1999. xiv, 283 pp.

46 **YAMAMOTO, Mutsumi:** Animacy and Reference. A cognitive approach to corpus linguistics. 1999. xviii, 278 pp.

45 **COLLINS, Peter C. and David LEE (eds.):** The Clause in English. In honour of Rodney Huddleston. 1999. xv, 342 pp.

44 **HANNAY, Mike and A. Machtelt BOLKESTEIN (eds.):** Functional Grammar and Verbal Interaction. 1998. xii, 304 pp.

43 **OLBERTZ, Hella, Kees HENGEVELD and Jesús SÁNCHEZ GARCÍA (eds.):** The Structure of the Lexicon in Functional Grammar. 1998. xii, 312 pp.

42 **DARNELL, Michael, Edith A. MORAVCSIK, Michael NOONAN, Frederick J. NEWMEYER and Kathleen M. WHEATLEY (eds.):** Functionalism and Formalism in Linguistics. Volume II: Case studies. 1999. vi, 407 pp.

41 **DARNELL, Michael, Edith A. MORAVCSIK, Michael NOONAN, Frederick J. NEWMEYER and Kathleen M. WHEATLEY (eds.):** Functionalism and Formalism in Linguistics. Volume I: General papers. 1999. vi, 486 pp.

40 **BIRNER, Betty J. and Gregory WARD:** Information Status and Noncanonical Word Order in English. 1998. xiv, 314 pp.

39 **WANNER, Leo (ed.):** Recent Trends in Meaning–Text Theory. 1997. xx, 202 pp.

38 **HACKING, Jane F.:** Coding the Hypothetical. A comparative typology of Russian and Macedonian conditionals. 1998. vi, 156 pp.

37 **HARVEY, Mark and Nicholas REID (eds.):** Nominal Classification in Aboriginal Australia. 1997. x, 296 pp.

36 **KAMIO, Akio (ed.):** Directions in Functional Linguistics. 1997. xiii, 259 pp.

35 **MATSUMOTO, Yoshiko:** Noun-Modifying Constructions in Japanese. A frame semantic approach. 1997. viii, 204 pp.

34 **HATAV, Galia:** The Semantics of Aspect and Modality. Evidence from English and Biblical Hebrew. 1997. x, 224 pp.

33 **VELÁZQUEZ-CASTILLO, Maura:** The Grammar of Possession. Inalienability, incorporation and possessor ascension in Guaraní. 1996. xvi, 274 pp.

32 **FRAJZYNGIER, Zygmunt:** Grammaticalization of the Complex Sentence. A case study in Chadic. 1996. xviii, 501 pp.

31 **WANNER, Leo (ed.):** Lexical Functions in Lexicography and Natural Language Processing. 1996. xx, 355 pp.

30 **HUFFMAN, Alan:** The Categories of Grammar. French lui and le. 1997. xiv, 379 pp.

29 **ENGBERG-PEDERSEN, Elisabeth, Michael FORTESCUE, Peter HARDER, Lars HELTOFT and Lisbeth Falster JAKOBSEN (eds.):** Content, Expression and Structure. Studies in Danish functional grammar. 1996. xvi, 510 pp.

28 **HERMAN, József (ed.):** Linguistic Studies on Latin. Selected papers from the 6th International Colloquium on Latin Linguistics (Budapest, 23–27 March 1991). 1994. ix, 421 pp.

27 **ABRAHAM, Werner, T. GIVÓN and Sandra A. THOMPSON (eds.):** Discourse, Grammar and Typology. Papers in honor of John W.M. Verhaar. 1995. xx, 352 pp.

26 **LIMA, Susan D., Roberta L. CORRIGAN and Gregory K. IVERSON:** The Reality of Linguistic Rules. 1994. xxiii, 480 pp.

25 **GODDARD, Cliff and Anna WIERZBICKA (eds.):** Semantic and Lexical Universals. Theory and empirical findings. 1994. viii, 510 pp.

24 **BHAT, D.N.S.:** The Adjectival Category. Criteria for differentiation and identification. 1994. xii, 295 pp.

23 **COMRIE, Bernard and Maria POLINSKY (eds.):** Causatives and Transitivity. 1993. x, 399 pp.

22 **McGREGOR, William B.:** A Functional Grammar of Gooniyandi. 1990. xx, 618 pp.

21 **COLEMAN, Robert (ed.):** New Studies in Latin Linguistics. Proceedings of the 4th International Colloquium on Latin Linguistics, Cambridge, April 1987. 1990. x, 480 pp.

20 **VERHAAR, John W.M. S.J. (ed.):** Melanesian Pidgin and Tok Pisin. Proceedings of the First International Conference on Pidgins and Creoles in Melanesia. 1990. xiv, 409 pp.

19 **BLUST, Robert A.:** Austronesian Root Theory. An essay on the limits of morphology. 1988. xi, 190 pp.

18 **WIERZBICKA, Anna:** The Semantics of Grammar. 1988. vii, 581 pp.

17 **CALBOLI, Gualtiero (ed.):** Subordination and Other Topics in Latin. Proceedings of the Third Colloquium on Latin Linguistics, Bologna, 1–5 April 1985. 1989. xxix, 691 pp.

16 **CONTE, Maria-Elisabeth, János Sánder PETÖFI and Emel SÖZER (eds.):** Text and Discourse Connectedness. Proceedings of the Conference on Connexity and Coherence, Urbino, July 16–21, 1984. 1989. xxiv, 584 pp.

15 **JUSTICE, David:** The Semantics of Form in Arabic. In the mirror of European languages. 1987. iv, 417 pp.

14 **BENSON, Morton, Evelyn BENSON and Robert F. ILSON:** Lexicographic Description of English. 1986. xiii, 275 pp.

13 **REESINK, Ger P.:** Structures and their Functions in Usan. 1987. xviii, 369 pp.

12 **PINKSTER, Harm (ed.):** Latin Linguistics and Linguistic Theory. Proceedings of the 1st International Colloquium on Latin Linguistics, Amsterdam, April 1981. 1983. xviii, 307 pp.

11 **PANHUIS, Dirk G.J.:** The Communicative Perspective in the Sentence. A study of Latin word order. 1982. viii, 172 pp.

10 **DRESSLER, Wolfgang U., Willi MAYERTHALER, Oswald PANAGL and Wolfgang Ullrich WURZEL:** Leitmotifs in Natural Morphology. 1988. ix, 168 pp.

9 **LANG, Ewald and John PHEBY:** The Semantics of Coordination. (English transl. by John Pheby from the German orig. ed. 'Semantik der koordinativen Verknüpfung', Berlin, 1977). 1984. 300 pp.

8 **BARTH, E.M. and J.L. MARTENS (eds.):** Argumentation: Approaches to Theory Formation. Containing the Contributions to the Groningen Conference on the Theory of Argumentation, October 1978. 1982. xviii, 333 pp.

7 **PARRET, Herman, Marina SBISÀ and Jef VERSCHUEREN (eds.):** Possibilities and Limitations of Pragmatics. Proceedings of the Conference on Pragmatics, Urbino, July 8–14, 1979. 1981. x, 854 pp.

6 **VAGO, Robert M. (ed.):** Issues in Vowel Harmony. Proceedings of the CUNY Linguistics Conference on Vowel Harmony, May 14, 1977. 1980. xx, 340 pp.

5 **HAIMAN, John:** Hua: A Papuan Language of the Eastern Highlands of New Guinea. 1980. iv, 550 pp.

4 **LLOYD, Albert L.:** Anatomy of the Verb. The Gothic Verb as a Model for a Unified Theory of Aspect, Actional Types, and Verbal Velocity. (Part I: Theory; Part II: Application). 1979. x, 351 pp.

3 **MALKIEL, Yakov:** From Particular to General Linguistics. Selected Essays 1965–1978. With an introduction by the author, an index rerum and an index nominum. 1983. xxii, 659 pp.

2 **ANWAR, Mohamed Sami:** BE and Equational Sentences in Egyptian Colloquial Arabic. 1979. vi, 128 pp.

1 **ABRAHAM, Werner (ed.):** Valence, Semantic Case, and Grammatical Relations. Workshop studies prepared for the 12th International Congress of Linguists, Vienna, August 29th to September 3rd, 1977. xiv, 729 pp. *Expected*
Out of print