IMPACT STUDIES IN LANGUAGE AND SOCIETY 28

Language Contact

New perspectives

EDITED BY

Muriel Norde
Bob de Jonge
Cornelius Hasselblatt

JOHN BENJAMINS PUBLISHING COMPANY

Language Contact

IMPACT: Studies in Language and Society

IMPACT publishes monographs, collective volumes, and text books on topics in sociolinguistics. The scope of the series is broad, with special emphasis on areas such as language planning and language policies; language conflict and language death; language standards and language change; dialectology; diglossia; discourse studies; language and social identity (gender, ethnicity, class, ideology); and history and methods of sociolinguistics.

Volume 28

Language Contact. New perspectives
Edited by Muriel Norde, Bob de Jonge and Cornelius Hasselblatt

Language Contact

New perspectives

Edited by

Muriel Norde
Bob de Jonge
Cornelius Hasselblatt
University of Groningen

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 contact : new perspectives / edited by Muriel Norde, Bob de Jonge and Cornelius Hasselblatt.
p. cm. (IMPACT: Studies in Language and Society, ISSN 1385-7908 ; v. 28)
Includes bibliographical references and index.
1. Languages in contact. I. Hasselblatt, Cornelius. II. Jonge, Bob de. III. Norde, Muriel, 1968-
P130.5.L336 2010
306.44--dc22 2009048317
ISBN 978 90 272 1867 4 (Hb ; alk. paper)
ISBN 978 90 272 8843 1 (Eb)

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

Acknowledgements

This volume contains a selection of papers presented at the *Language Contact in Times of Globalization* conference, held at the University of Groningen, September 28–30, 2006, organized by the editors of this volume. We gratefully acknowledge the financial support of *The Netherlands Organization for Scientific Research* (NWO), *The Royal Netherlands Academy of Arts and Sciences* (KNAW), the *Center for Language and Cognition Groningen* (CLCG), and the *Stichting Groninger Universiteitsfonds* (GUF). We also wish to thank all the participants for making the conference such a success, and our indispensable student assistants who took care of all things practical: Martijn Boonstra, Pieter Goossens, and Rosine Scheirs. Special thanks are due to the referees who commented on the first selection of papers: Uwe Bläsing, Charlotte Gooskens, Björn Hansen, Peter Houtzagers, Frederik Kortlandt, Stella Linn, Virve Raag, Gisela Redeker, Anneli Sarhimaa, Willem Visser, and Ludger Zeevaert.

Introduction

Cornelius Hasselblatt, Bob de Jonge and Muriel Norde
University of Groningen

The year 2008 marked the 20th anniversary of Thomason & Kaufman's *Language Contact, Creolization, and Genetic Linguistics*. This work, a convincing combination of a theoretical framework and detailed case studies, has given a tremendous boost to the study of language contact phenomena, and it still stands as the most influential work in the field since the publication of Weinreich's (1953) foundational *Languages in Contact*.

In the years following the publication of Thomason & Kaufman's book, parallel political and technical developments – the end of the Cold War and the internet revolution – gave an additional impetus to language contact research in many respects: an increasing number of languages and data became (electronically) accessible, the mobility of people grew, and new contact situations came into being. Indeed, even "new" languages arose (e.g. Croatian and Serbian, where the Slavists of one generation earlier would have spoken of Serbo-Croat). Today no-one seriously doubts the impact language contact had on (probably) any of the world's languages. Opinions only differ with respect to the scope and intensity of it.

Research into language contact can be roughly divided into three main branches. The first is the traditional in-depth investigation of a certain contact situation, usually involving no more than two languages. This has been one of the goals of historical comparative linguistics from its very beginning and will remain important in the near future, too, as the field is still far from being completely explored. Countless parts of the world and contact situations remain to be examined. Even long-standing and well-known contacts had not been investigated exhaustively until recently (e.g. Sarhimaa 1999 on Russian-Karelian contact, De Smit 2006 on Finnish-Swedish contact, Braunmüller & Diercks 1993 and Braunmüller 1995 on Low German-Scandinavian contact in the Late Middle Ages, or Silva-Corvalán 1994 on language contact between Spanish and English in Los Angeles). A second group of scholars has been focusing on cross-linguistic comparisons and the identification of larger linguistic areas (cf. e.g. Heine & Kuteva 2005, Heine & Kuteva 2006, but already Décsy 1973). A third major topic has been the negative effects of intensive contact, with language or dialect death as a possible result (cf. e.g. Nettle & Romaine 2000, Janse & Tol 2003).

The papers in this volume are concerned with different levels and different aspects of language contact in a wide variety of languages and language families, including Indo-European (Germanic, Romance, and Slavic), Finno-Ugric, Turkic, and Japanese.

Pieter **Muysken's** paper is a theoretical contribution to ethnolect research. Muysken discusses two different views on ethnolects: the "shift perspective" and the "multidimensional perspective". In the current literature, the shift perspective has attracted most attention; within this perspective, focus is on the approximation of the ethnic group towards the national language and allows comparisons between ethnic varieties and standard varieties of the dominant language in a speech community. In the multidimensional perspective, however, the original languages of the ethnic group as well as processes of mutual convergence and simplification play an additional role in the new varieties. Moreover, this perspective allows us to contextualize the varieties within an overall account of the multilingual repertoires of speakers of a non-dominant language and of the strategies that they employ to make use of this repertoire.

The next paper is a case study in shifts effects in ethnolects. In this paper, John **Nerbonne**, Timo **Lauttamus**, Lisa Lena **Opas-Hänninen**, and Wybo **Wiersma** explore techniques to automatically tag corpora and to detect syntactic differences between first generation and second generation Finnish immigrants in Australia. On the basis of a corpus of interviews comprising some 305,000 tokens, they found that first generation speakers were significantly more prone to 'syntactic contamination' from Finnish. The authors were also able to identify specific syntactic phenomena in the speech of first generation speakers, such as omission of progressive auxiliary 'be', existential 'there', and anaphoric 'it'. They conclude that some of the features found in the data can certainly be explained as contact-induced changes, whereas others may be ascribed to universally determined properties of the language faculty.

Where Muysken's and Nerbonne et al's papers are concerned with varieties of the national language spoken by ethnic groups, the subsequent three papers discuss the reverse: changes in immigrant languages resulting from contact with the dominant national language. In the first of these, Ricardo **Otheguy**, Ana Celia **Zentella** and David **Livert** consider a number of grammatical factors that can be held responsible for attested variation in the usage of subject pronouns by Spanish-speaking immigrants in New York City. The authors conclude that the increase in subject pronoun rates in the immigrants in NYC is due to their adapting their usage of Spanish pronouns to that of the equivalent pronouns in English. This observation is corroborated by various factors, such as a correlation with the number of years that they have lived in NYC, their age of arrival, their English skills, and different settings in which the languages are used.

Piibi-Kai **Kivik's** paper likewise deals with pronoun variation in an immigrant language in the United States, viz. Estonian. Kivik's study contrasts the use of personal pronouns in Estonian spoken by first-generation immigrants or long-term sojourners in the United States, and the use of pronouns by monolingual speakers in Estonia. She found that the pronoun use in varieties influenced by English differs from that of monolinguals living in Estonia. For example, bilingual Estonian-English speakers make less use of zero subjects, which are a feature of standard Estonian but not of English, and can therefore clearly be classified as contact-induced change.

The third and final paper on changes in immigrant languages is the contribution by A. Seza **Doğruöz** and Ad **Backus**. Their purpose is to investigate whether there exist unconventional constructions in the Turkish spoken by Turkish immigrants in the Netherlands, and if so, if these could be attributed to Dutch influence. At a superficial glance, both questions can be answered positively, but the authors found that the observed unconventional constructions also existed in standard Turkish. This may be taken as proof that language change in situations of language contact accelerates incipient changes that already existed in the standard language (Silva-Corvalán 1986: 588, 604).

A historical perspective on the effects of language contact resulting from immigration is provided in a paper by Charlotte **Gooskens**, Renée **van Bezooijen** and Sebastian **Kürschner.** This paper is concerned with loan words in the national language from a culturally dominant immigrant language. The authors present a corpus-based survey of the percentage of loan-words in Dutch and Swedish, arguing that the differences between Dutch and Swedish can be explained both by linguistic distance and type of contact situation. Drawing data from the *Europarl* corpus of speeches held in the European Parliament and their translations, from which they extracted the most frequent 15,000 words for both languages, they show that the percentage of loans is significantly higher in Swedish (44.4%) than in Dutch (27.9%). One particularly striking difference regards the percentage of Low German loans, which is much higher in Swedish, even though both Swedish and Dutch had profound contacts with Low German speaking merchants in the Hanseatic period. This is due, the authors argue, to the fact that Dutch and Low German belong to one dialect continuum, which makes it very difficult to distinguish between native Dutch words and Low German loans. Gooskens et al's findings are in accordance with previous (small-scale) studies in the impact of foreign languages on the Swedish and Dutch lexicons.

Needless to say, immigration is not the only situation which may give rise to contact-induced change – border areas form another domain where contact effects are to be expected. Two of the papers in this volume deal with language contact in such areas. The first, by Hélène **Brijnen**, analyses the use of *gor*, a particle

borrowed from German (*gar*), in the Eastern Sorbian border dialect. Among other things she found that the Upper Sorbian dialects with strong Lower Sorbian influence show a more frequent use of *gor* than those without Lower Sorbian influence. According to the author the usage of *gor* is constrained by the phonological system of the recipient dialect – *gor* is more frequent in varieties which possess the consonant [g] than in varieties which lack this consonant.

Wilbert **Heeringa**, John **Nerbonne** and Petya **Osenova's** paper is a computational linguistic approach to border area contacts, exploring techniques that can be used to measure the effects of language contact. For this purpose they compared a selection of Bulgarian dialects to the five neighbouring languages Macedonian, Serbian, Romanian, Greek and Turkish, hypothesizing that dialects in the vicinity of one of these languages are phonologically more similar to these languages than dialects spoken further away. In order to operationalize the notion of "phonologically similar", the authors applied three different techniques: "Levenshtein distances", which aligns corresponding segments of pairs of (preferably cognate) words and sums the differences between these segments; the "phone frequency method", which compares two languages or language varieties by counting the number of tokens of each phoneme in comparable corpora; and the "feature frequency method", which counts the number of tokens of segments with specific values for given phonological features. All three techniques detected positive correlations between geographic and phonological distances in the case of Macedonian, Serbian and Romanian, which is a remarkable result in the light of the traditional assumption that phonology is only marginally affected in Balkan *Sprachbund* contacts. Another surprising effect was that the computational analysis showed negative correlations for Greek and Turkish, which the authors tentatively assume to be the consequence of historical and /or sociolinguistic factors. They finally found that the three techniques do not always correlate well with each other, which provides an important direction for future research.

Finally, languages may change without physical contact with speakers of another language. English is an obvious example of a language affecting other languages without direct contacts between speakers. Jason **Shaw** and Rahul **Balusu**'s paper concerns the introduction of the phonological contrast between /t/ and /tʃ/ before high front vowels in present-day Japanese. Data were gathered by means of an oral elicitation test in which two generations of Japanese native speakers, with little or no conversational proficiency in English, participated. The results of the test show that the older generation produce a weak contrast between /ti/ and /tʃi/ in some loans, by mapping these sequences to prosodically conditioned allophones of native /ti/. This contrast was subsequently enhanced by the

younger generation. Since most of the English loans date from after 1975, the older generation must have acquired this contrast during adulthood, suggesting that phonological change is possible even after the "critical period" of language acquisition. The authors furthermore show that the borrowing of phonological contrasts is constrained by existent allophonic variation in the target language.

Translators play an obvious role in long-distance language contact. In the final contribution to this volume, Nicola **Borrelli** investigates to what extent the translations of Brussels' official documents mirror the specific national perspectives of their translators, showing that they cannot escape the influence of the public opinion of their home countries.

The papers presented in this volume show a great variety in sources and methods, but they all contribute to our knowledge and understanding of language contact in times of globalization, while at the same time offering suggestions for further research.

References

Braunmüller, K. (ed.). 1995. *Niederdeutsch und die skandinavischen Sprachen* II. Heidelberg: C. Winter.

Braunmüller, K. & Diercks, W. (eds). 1993. *Niederdeutsch und die skandinavischen Sprachen* I. Heidelberg: C. Winter.

Décsy, G. 1973. *Die linguistische Struktur Europas. Vergangenheit – Gegenwart – Zukunft*. Wiesbaden: Harrassowitz.

Heine, B. & Kuteva, T. 2005. *Language Contact and Grammatical Change*. Cambridge: CUP.

Heine, B. & Kuteva, T. 2006. *The Changing Languages of Europe*. Oxford: OUP.

Janse, M. & Tol, S. (eds). 2003. *Language Death and Language Maintenance. Theoretical, Practical and Descriptive Approaches* [Amsterdam Studies in the Theory and History of Linguistic Science. Series IV – Current Issues in Linguistic Theory 240]. Amsterdam: John Benjamins.

Nettle, D. & Romaine, S. 2000. *Vanishing voices. The extinction of the world's languages*. Oxford: Oxford University Press

Sarhimaa, A. 1999. *Syntactic Transfer, Contact-induced Change, and the Evolution of Bilingual Mixed Codes. Focus on Karelian-Russian Language Alternation* [Studia Fennica Linguistica 9]. Helsinki: Finnish Literature Society.

Silva-Corvalán, C. 1986. Bilingualism and language change. *Language* 62: 587-608.

Silva-Corvalán, C. 1994. *Language Contact and Language Change. Spanish in Los Angeles*. Oxford: Clarendon Press

Smit, M. de. 2006. *Language Contact and Structural Change. An Old Finnish Case Study* [Acta Universitatis Stockholmiensis. Studia Fennica Stockholmiensia 9]. Stockholm.

Thomason, S. G. & Kaufman, T. 1988. *Language Contact, Creolization, and Genetic Linguistics*. Berkeley CA: University of California Press.

Weinreich, U. 1953[1964]. *Languages in Contact. Findings and Problems* [Publications of the Linguistic Circle of New York 1]. The Hague: Mouton.

Ethnolects as a multidimensional phenomenon

Pieter Muysken
Radboud Universiteit Nijmegen

This paper contrasts two different views of the phenomenon of ethnolect, ethnic varieties of a language: the *shift* perspective and the *multidimensional* perspective. In the shift perspective, the focus is on the approximation in the speech of ethnic groups to the dominant national target language, while in the multidimensional perspective the original languages of the ethnic group and processes of mutual convergence and simplification play an additional role. The multidimensional perspective allows us to contextualize the varieties that have emerged in the process of shift within an overall account of multilingual repertoires of speakers of a non-dominant language. The complexity of the verbal repertoire and the alternating reliance on strategies of maintenance and shift, convergence, mixing, and simplification create the ethnolectal varieties.

1. Introduction[1]

Migrations of large groups of speakers to another country, or to metropolitan regions in their own country where a different language is dominant, have produced notable changes in the language behavior of these speakers. Similarly, smaller ethnic groups in a country where a language different from their own is dominant are under increasing pressure to adapt to this dominant language, under the aegis of national integration. Migration and national integration, two facets of the phenomenon of globalization, have conjointly led to the emergence of ethnic varieties: ethnolects.

---

1. The research reported on here is part of a group project *Roots of Ethnolects* funded by the Netherlands Organization for Scientific research NWO with the participation of Hanke van Buren, Frans Hinskens, Ariën van Wijngaarden, and myself. In an early stage the project also involved Esther Krieken and Wouter Kusters. The perspectives taken here do not necessarily reflect the views of my colleagues in the project.

This paper contrasts two different views of the phenomenon of ethnolect, views which may be labeled the *shift* perspective and the *multidimensional* perspective. In the shift perspective, the focus is on the approximation in the speech of ethnic groups to the dominant national target language, while in the multidimensional perspective the original languages of the ethnic group as well as processes of mutual convergence and simplification play an additional role. The shift perspective is dominant in the current literature, but the complementary multidimensional perspective is also an important one, and keeps creeping up. In other words, we can speak of a specific variety (the shift perspective) versus the community repertoire (the multidimensional perspective).

The study of ethnolects has much more than a purely academic interest since it is heavily loaded in terms of the notions of national identity, language purity, and language diversity. To take the example of Dutch in the Netherlands (but the same goes for many other languages of industrial nations): there is a serious concern about the level of the proficiency in Dutch in e.g. school populations, not just among students of an immigrant background. Some people think that the level of control of Dutch should be the object of concern and intervention; this is a view shared by teachers and policy makers alike. The limited knowledge of Dutch, it is often the implicit message of popular discourse, would spread from young people from immigrant families to those from established family backgrounds. In this perspective, ethnolects are often equated with language decay. The continued use of the original languages of minority groups, migrant or historically rooted in the country, adds to these concerns.

I do not really know whether we can speak of decay of the national languages in the schools. Personally, I think that today's students have different skills – at least in part – from those of twenty years ago, which makes comparisons difficult. Intriguing is the possibility that the nature of a language such as Dutch is changing under the influence of migrant populations acquiring it as a second language, as in the shift perspective. When comparing native and non-native students we also need to take account of social class. I.e. the majority of non-native students are working class and concerns about linguistic skills in this social group have been voiced long ago, cf. Bernstein's work on the restricted code (1966).

2. The shift perspective

All over the world ethnic varieties of larger national languages are emerging, and of course have emerged for many centuries, as the result of second language acquisition after migration or integration into a larger political entity. When large groups of people thus acquire a second language, often in a process of language

shift, the language is subtly transformed and the result is then called an ethnolectal variety. For this reason, research on these shift ethnolects has become an increasingly important area within language contact studies. The process of group shift has occurred both in migration settings and in contexts where a resident minority population has shifted to a dominant national language. Since both migration and integration into larger national units, have triggered counter-reactions in terms of the reaffirmation of ethnic and regional identities, group language shift is an increasingly important area of study. The study of ethnolects in the shift perspective is thus concerned with the more or less stable outcomes of this process of group shift, particularly in urban settings (Wölck 1984).[2]

In the corresponding definition, an ethnolect is "the variety of a language that results when speakers of different ethnolinguistic backgrounds attempt to speak the dominant language (e.g. 'Chicano English')" (Danesi 1985: 118). In Danesi's conception, ethnolects are thus products of language shift in the sense of Thomason & Kaufman (1988).

In a similar vein, Androutsopoulos (2001: 2) defines an ethnolect as "a variety of the majority language (or 'host language') which is used and regarded as a vernacular for speakers of a particular ethnic descent and is marked by certain contact phenomena".

3. The multidimensional perspective

According to some researchers, however, ethnolects are not strictly to be equated with learners' varieties. In some communities, we find maintenance of a minority language. In communities where immigration took place a long time ago, such as the ethnic Italians or Polish in North America, most speakers are proficient in the standard variety. In the case of these speakers the resulting ethnolect is not a matter of not being able to, but rather of (under certain conditions) not wanting

2. The project *Roots of Ethnolects* is primarily focused on the shift perspective on ethnolects. We record young people in two cities, Amsterdam and Nijmegen, to determine the influence of the regional pronunciation on the ethnolect. Indeed Moroccan young people sound very different in both cities. Amsterdam and Nijmegen non-standard Dutch play a role, but also the own group languages of the young people, in our case Berber and Turkish. Furthermore we compare 11 and 19 year olds to establish the role of further language development, and we record paired conversations of young people in little groups with different partners (Moroccan, Dutch, Turkish) to determine the role of the background of the conversational partner, and whether the ethnolects influence the language use of young people of other ethnic backgrounds. Finally we hope to gain insight into the longer term effects on the Dutch language in an urban context of the presence of young people with so many different backgrounds.

to speak standard or, more generally, to sound more native. In yet other communities, such as the Mexican immigrants in California and Texas, we find extensive Spanish-English language mixing. Accordingly, definitions of the notion of ethnolect actually vary quite a bit in the literature. Danesi in an earlier publication (1984: 100) gives a definition which focuses on the original language of the speech community: '… *ethnolect* of the mother tongue. This can be defined generally to be a version of the language of origin which, primarily as a consequence of the frequent borrowing and adoption of words from the culturally dominant language, has come to characterize the speech habits of the immigrant community.' Clyne (2000: 86) gives a more neutral definition, describing ethnolects as 'varieties of a language that mark speakers as members of ethnic groups who originally used another language or distinctive variety'. Below I will further explore the view that ethnolects may in fact be multi-dimensional, multi-faceted phenomenon.

To illustrate this, consider again the example of Mexican Americans. Fought (2006: 74–5) outlines the different varieties or codes to be found in the verbal repertoires of different Latino communities in the U.S., ranging from 'Standard' English to 'Standard' Spanish:

'Standard' English
Other local varieties of English (e.g. from the South in the U.S.)
Afro-American Vernacular English (AAVE)
Varieties of Latino English in different Latino communities
Non-native Spanish-influenced English
Code-switching and -mixing
Non-native English-influenced Spanish
Varieties of Latino Spanish
Regional varieties of Spanish (e.g. Puerto-Rican Spanish)
'Standard' Spanish

Members of the Latino community will generally not command this full range, but certainly a fair chunk of it, depending on their regional and social background, and their language practices in daily life. The middle parts of this range, including 'Varieties of Latino English in different Latino communities', 'Non-native Spanish-influenced English', 'Code-switching and -mixing', 'Non-native English-influenced Spanish', and 'Varieties of Latino Spanish' could well be included under the label 'Latino ethnolect'. The two types of English in this range are linked to each other, as are the two types of Spanish.

The English and Spanish varieties meet each other in code-switching and -mixing. 'In most Latino communities, code-switching plays a crucial role,' according to Fought (2006: 75). The resulting varieties have various regional labels: Spanglish (Puerto-Ricans in New York), Tex-Mex (Texas), and Pachuco or Caló

(Chicano) in the Southwestern United States. In many popular accounts, these are truly separate and distinct varieties in their own right (often looked down upon), while many linguists will stress the fact that they are the product of combining two distinct languages (each of which may have undergone numerous influences from the other one, to be sure). In the code-switching literature, evidence for both perspectives can be found. On the one hand, people who switch a lot can also produce sentences in either language without switching, depending on their interlocutor – evidence for the two separate varieties account. On the other, switching often involves special linguistic mechanisms, patterns, or constructions which are not found in either language – evidence for the single mixed variety account.

In a dynamic and multi-dimensional perspective, these various dimensions and possibilities can be integrated.

4. Dimensions of a definition

Departing from the external definitions of ethnolects as specific varieties of a language spoken by a recognizable and delineable ethnic group that is part of a larger community, the next question is what features they have and how they emerge. Key elements in the answers to these questions are obviously second language acquisition, maintaining a separate identity, and mixing or transfer.

The general public is not familiar with the notion of ethnolect, and even among linguists this is not a household word. Primary reactions among beginning university language students in the Netherlands, when you ask them what an ethnolect is, are typically the following:

a. Ethnolects are simply incompletely acquired Dutch;
b. Ethnolects are street language, symbols of young peoples' identity;
c. Ethnolects are mixed language, half Turkish (or Moroccan), half Dutch;
d. Ethnolects are simply the local variety of the immigrant languages.

I will argue that all four cases involve something that has to do with ethnolects, but that in every instance we are dealing with another phenomenon, which contributes a dimension to the notion of ethnolect. What are the features then of language acquisition, the learning of a new language, of street language, and of mixed language or code switching, so that they on the one hand should be distinguished from ethnolects, and on the other hand feed into them? I will describe them in terms of a number of features, and subsequently I will show how ethnolects are distinct from but also tie up with these different other forms of language use. Unless indicated otherwise the italicized cited sentences come from material that we have collected so far in the research project called *Roots of Ethnolects*, discussed below.

4.1 Language acquisition

When a Moroccan child, here an 11-year old born in the Netherlands, says something like (1) or (2), this is clearly a sign of lack of proficiency in the Dutch language:

(1) *Hij heeft een bril aan.*
He has [a] glasses at.
'He's wearing glasses'

(2) *Iedereen draagt die kleren aan.*
Everyone wears those clothes at.
'Everyone's wearing those clothes'

In the case of (1) the preposition *aan* 'at' is used instead of *op* 'on', and in (2) *aan* is used with a verb that does not call for that preposition generally. There is a vast literature on second language acquisition, also where Dutch as the target language is concerned, which can not be summarized here (cf. e.g. Van de Craats 2000; Starren, 2001).

This type of phenomena is generally not stable: acquisition errors are sometimes made by the child, but not at other times. Often these are very individual errors though: something this child has created, and another child makes yet again another type of errors. There are also acquisition problems that more children have, however, a shared acquisition pattern, due to 'universal' developmental tendencies. The type of errors in itself is not a signal of identity: the child would prefer not to make the error. Acquisition errors may be characteristic of a particular language group (for example in the order of the constituents), but are often characteristic of individuals from all sorts of language groups. Neither is this type of errors really tied to a particular setting: the child could make that error in all circumstances. All words are Dutch, and correspond to the ordinary vocabulary of Dutch. The L2 utterance may show special pronunciation or sentence intonation, and deviating grammatical basic patterns, but to a different extent for different learners. There are also deviating endings (for example in the finite verb), and cases of periphrasis instead of endings of complex words.

Concepts are often paraphrased in the L2 variety, as for example in (3):

(3) *Meester, moet dat ding per se ons ding nadenken?*
Teacher, must that thing really our thing think.about?
'Teacher, must that thing really think about our thing?'

The child means to say something like (4)

(4) *Meester, neemt dat apparaat onze taal per se op?*
Teacher, records that machine our language really on?
'Teacher, is that machine really recording our language?'

Here *apparaat* 'machine' is paraphrased as *dat ding* 'that thing', and *onze taal opnemen* 'our speech record' by *ons ding nadenken* 'our thing think'. Notice also that a modal moet 'must' is used in (3), while the context does not require a modal.

Acquisition features can enter into the ethnolect, but many are not systematic enough to be adopted by the ethnolect, nor are they tied to an individual or a moment of speech.

4.2 Street language

Street language, also named young people's language, and Murks in Utrecht (Nortier 2001), is used by young people from all kinds of ethnic backgrounds, especially in the cities (Kotsinas 1998). Characteristic of street language in the Netherlands is the large number of words from different languages, mostly Sranan, and in addition Moroccan Arabic (especially in Utrecht but also elsewhere) and Cape Verdean (especially in Rotterdam). Young people with an originally Dutch background use street language too, but the special words do not all come from the ethnic group of the person that uses street language at that moment, although this can be the case, to be sure... Apart from words from a specific ethnic group words from English or Spanish may be used, as well as newly made words and abbreviations. Street language is not stable; it can differ considerably from person to person, place to place, situation to situation, but it is not strictly individual; it is a shared code. It is a signal of identity, but of a collective young people's identity rather than of a specific ethnic identity. At the same time it is tied to a situation, a way of speaking with particular people, at a specific moment. It is clearly recognizable for many people, even though adult outsiders differ widely in their knowledge of street language words. Many Dutch people will know what *dissen* (to show disrespect for someone) and *chill* (cool) mean, but have no idea what is meant by *doekoe* (money) or *faja* (bothersome), and that a *tori* is a story. All three words are from the Surinam Creole language Sranan, one of the primary lexical sources for street language in the Netherlands.

Some young people also have a special pronunciation and sentence intonation when they speak street language, but that is not an essential feature. Sometimes Dutch adjectives get the ending *–e* where they should not, a form of overgeneralization: a *leuke meisje* (a nice girl), but finite verbs always receive the correct ending (in any case following the norms of ordinary spoken language, so *hij heb* for 'he has' rather than standard *hij heeft*). For the rest the rules of grammar are applied correctly, and there are no clear instances of deviating word order or the like.

4.3 Mixed language

Groups of young people of Turkish or Surinamese backgrounds speaking among each other often switch between their languages. More generally, some bilinguals mix their two languages very strongly, inside of a conversation, inside of an utterance, and sometimes even inside of a sentence constituent. This phenomenon has been thoroughly investigated: the best studies of Moroccan Arabic-Dutch language mixing are those of Jacomien Nortier (1990) and Louis Boumans (1999), while Ad Backus (1996) has carried out very extensive research on the Turkish mixed language. In addition work has been done on mixed language use involving Surinam Creole (Bolle 1994) and Hindustani (Kishna 1979) in the Surinamese community, on Moluccan code switching (Huwae 1992), and on Papiamento-Dutch code-switching (Muysken, Kook and Vedder 1996). The overall results are more or less the same: in particular young people are very well capable of using their languages in a mixed way, inside of one clause, but in addition are also able to keep the languages separate. The following examples come from the material of Backus (1992; cf. also Backus 1996) and represent Turkish-Dutch language mixing among the intermediate generation, born in Turkey but moved to the Netherlands at an early age (Dutch fragments are in bold).

(5) ***Cassetterecorder****-nan* ***frietten****t-e* *gidelim* *la*
tape.recorder-with fries.booth-to go-wish-we Q
'Shall we go with the tape recorder to the fries booth?'

(6) *bir tane* ***donkere jongen****-nan*
one piece dark boy-with
'with a dark boy'

(7) ***Engels****-I* *bir tane* ***blonde meisje****-den* *alIyordun*
English-AC one piece blond girl-from you got
'You got English from a blond girl.'

This type of language use is naturally very much tied to a peer group, and requires a good knowledge of the original language of the group of immigrants. It is clearly also a way of talking that is only possible with young people from the group itself, given that knowledge of Turkish is required.

4.4 Ethnolects as transplanted varieties

Finally, there is the view of ethnolect as the local transplanted variety of a minority migrant language, as in the Toronto Italian described by Danesi (1984) or the Turkish spoken in the Netherlands as analyzed by Doğruoz (2007). The extent to

which languages change when transplanted to new immigrant communities elsewhere is not clear, but has been studied considerably since Haugen's (1956) pioneering work on the Norwegians in the United States. The study of transplanted varieties has been particularly productive in the United States and in Australia, where Clyne (1967) initiated this line of research with his work on the Dutch and German immigrant communities. In Europe similar research emerged out of the study of various languages of immigrant communities in their wider context, such as Johanson's (1992) work on regular structural features of Turkish in contact with other languages; this type of research has also been taken up with respect to Turkish immigrant communities by a number of researchers in Germany, Norway, and the Netherlands, most recently the already cited Doğruoz (2007). There has also been work on Arabic in the Maghreb in North Africa and in different European countries. Boumans (2006) is a recent example; he studies the evolution of both synthetic and analytic attributive possession constructions in spoken Arabic in Morocco and the Netherlands, showing that there is an interaction of inherent tendencies of change and language contact factors.

Considering the diagnostic features of the different varieties that can be linked to ethnolects, it is clear that transplanted immigrant languages are more or less stable community languages, not individual creations, strong carriers of social and ethnic identity, and of course group specific, though not linked to specific settings. They obviously require thorough knowledge of the original community language, and are primarily based on home language words and expressions. They contain newly formed words, but not really many words from other language groups, except for the dominant national language. They may show particular patterns of pronunciation and sentence intonation, and may be influenced by non-standard grammatical and semantic patterns from the national language, as in *gitar oynamak* 'play the guitar' (following the Dutch model *gitaar spelen*), rather than the traditional form *gitar çalmak* 'sound the guitar' (Aarssen and Backus 2001: 318). However, the usage of inflections from the national language is rare. There may be periphrastic expressions replacing verb morphology, particularly with borrowed verbs, as in (8) (cited from Aarssen and Backus 2001: 319):

(8) *millet* ***kijken*** *yapıyor*
people watch.INF do.they
'People watch [you].'

In (8) the Dutch verb *kijken* 'watch' is used with a Turkish auxiliary verb *yap-* 'do'.

Schaufeli (1991), like Doğruoz (2007), found relatively few changes in Turkish in the Netherlands. One of phenomena found is the increased use of analytic forms for marking reported speech, a feature of discourse organization.

Table 1. Properties of the different dimensions of ethnolects and contributing varieties

	Ethnolect	L2 acquisition	Street language	Mixed language	(Transplanted) immigrant language
Stable	+	−	−	±	+
Individual	−	+	±	±	±
Signal of identity	+	−	+	+	+
Specific to a particular ethnic group	+	±	−	+	+
Specific to a particular setting	−	−	+	+	−
Requires a thorough knowledge of the original language of the own group	−	±	±	+	+
Words or constituents from the home language	−	−	±	+	+
Newly formed words	−	−	+	−	+
Words from the language of other group(s)	−	−	+	−	−
Particular pronunciation or sentence intonation	+	+	±	±	±
Non-standard grammatical basic patterns from the national language	±	+	−	+	±
Non-standard use of endings from the national language	+	+	±	±	−
Periphrasis instead of verb endings	±	+	+	−	±

4.5 Ethnolects as compared to these other phenomena

Ethnolects resemble the variants named so far closely, but have specific combinations of diagnostic property specifications of their own.

An ethnolect is a rather stable version of the dominant national language, and not tied to an individual but to a particular group (e.g. third generation Moroccan Dutch people). Although it will not always be used it is not specific to a particular setting. Words or phrases from the home language of the speakers, newly made words, or words from the languages of other group(s) occur infrequently. Table 1 gives an overview of all distinctions that I have made so far, for ethnolect, language acquisition, street language, and mixed language.

It becomes clear that the pluses and minuses in the different columns do not really correspond to each other. No single variety of the four listed corresponds directly to the idealized 'shift'-type ethnolect in the first column, although all share characteristics with it to some extent. Out of a total of 13 possible correspondence points (all pluses and minuses shared), the other varieties have beteen 5.5 and 8 values in common (where a –/± or +/± correspondence is counted as .5).

Ethnolects have a particular pronunciation or sentence intonation. This makes for example Surinamese Dutch immediately recognizable. Whether there are also non-standard grammatical basic patterns we do not really know in the case of all ethnolects involving Dutch; in the case of Surinam Dutch in any case there are many. Deviating use of endings, specifically the generalization of the adjective ending *–e* as in *een heel goede restaurant* ('a very good restaurant', should be *een heel goed restaurant*), is very often present. Periphrasis instead of endings of complex words are less often encountered. The use of *gaan* as a future auxiliary is striking (Cornips 2000), as in *Ik ga je slaan* 'I'll hit you.'

4.6 The perspective of the linguistic components involved

Already in the previous discussion a number of separate components or modules of the language system were alluded to. Here I attempt a more systematic overview, presented in Table 2.

A wide variety of components is involved in different dimensions of ethnolectal varieties, so far with a focus on aspects of 'outer' form. Different dimensions of the ethnolect, needless to say, tap into different grammatical components affected. This issue requires further exploration, however.

5. Towards an integrative perspective

Two views of ethnolects were presented above: a shift perspective, focusing on a particular variety of Dutch encountered, and the multidimensional perspective, focusing on the range of communicative possibilities of immigrants. However, these two perspectives are not so separate as suggested above. Ethnolects are fed by the other phenomena discussed: L2 acquisition, language mixing, transplanted language varieties, and street language, and would not be possible without these.

A case in point is L2 acquisition. In all of the following four sentences, a functional element is omitted. This occurs frequently with children with a low proficiency in Dutch, and at the same time we find similar examples also in the in-group language use of Moroccan and Turkish young people who actually speak Dutch fluently.

Table 2. Overview of the role of a number of separate components or modules of the language system in ethnolects

Component	Comments	Key reference
Sentence intonation	Very frequently mentioned	Carlock (1978)
Phrasal prosody	Particularly studied so far in relation to clausal organization	Selting (2006), (2007)
Segmental phonology	Studied for a number of ethnolects; some features such as stops for interdental fricatives are characteristic of several ethnolectal varieties of English	Fought (2003), (2006)
Content lexicon	Frequent L1 usage in street language	Kotsinas (1998)
Idioms and collocations	L1>L2 and L2>L1 influence	Vertschik (2007); Doğruoz (2007)
Morpho-syntactic distinctions	Frequent locus of variation. The morpho-syntax of ethnic varieties has been studied in particular detail with respect to African American Vernacular English	Current research of Ariën van Wijngaarden
Syntax and constituent order	Less affected	Doğruoz (2007)
Discourse markers and interjections	Frequent L1 usage in L2 shift varieties	Matras (2000); Dubois and Horvath (2002)

(9) *Ja, [het] is goed.*
Yes, [it] is ok.

(10) *[Een] Interview toch was het?*
[An] interview still was it?
'An interview, wasn't it?'

(11) *Maar dan is [het] voor mij ook [de] tweede keer hoor.*
But then is [it] for me also [the] second time hear.
But then [it] is for me, also [the] second time you know.

(12) *Maar nu [is hij] trainer.*
But now [is he] trainer.
But now [he is] a trainer.

Impersonal *het* 'it', indefinite (*een*) and definite (*de*) articles, the copula (*is*), and sometimes other pronouns are frequently omitted. Processes of language acquisition feed the ethnolect, and often the particular characteristics of the shift type ethnolect derive from language acquisition processes.

Consider chains of reference tracking, the ways in which speakers make clear what the entities are they are referring to, often with pronouns or anaphoric expressions, in the context of the discourse. Consider the following dialogue between Omer (Turkish) and Youssef (Moroccan), in which the rules for a particular card

game are discussed, from our recordings in the Roots project. Again, we notice frequent omissions, particularly also of the locative pronoun *er* (glossed R and corresponding to English *there* as in *thereof*):

(13) a. *Jij krijg-t [er] vier. Ik krijg [er] vier.* [Omer]
you get-2s [R] four. I get [R] four.
b. *Als je een negen heb-t, moet je negen gooi-en, en die zijn voor jou.*
If you a nine have-2s, must you nine throw-INF, and those are for you.
c. *Als je [die] niet heb-t, moet je wat anders gooi-en.*
If you [that] not have-2s, must you something else throw-INF
d. *En als ik, wat jij gooi-t, dezelfde wel heb, dan is [het] van mij.*
and if I, what you throw-2s, the.same have, then is [it] of me
e. *Je hebt nou een pisti. Deze moet je open leggen. Deze moet je der-op sluit-en.*
you have-2s now a pisti this must you open lay-INF this must you R-on close-INF
f. *Mag ik deze ook open draai-en of moet ik [het] zo lat-en?* [Youssef]
may I this also open turn-INF or must I [it] thus leave-INF
g. *Nee nee nee nee. Je moet [het] dicht draai-en.* [Omer]
no no no no you must [it] closed turn-INF
h. *die … die wat je heb gepakt daar-op.*
that … that what you have pick.PART R-on

Omer: 'You get four of them. I get four of them. If you have a nine, you must throw nine, and those are for you. If you do not have that, you must throw something else, and if I have the.same as what you throw, then it is mine. You now have a pisti. This you must lay open. This you must close on it.'
Youssef: 'May I also turn this open or must I leave it like this?'
Omer: 'No no no no you must [it] closed turn that … that what you have picked on.'

For language mixing the situation is a bit more complicated. With Moroccans few if any elements from Arabic or Berber enter into the ethnolect, with the exception of some interjections and exclamatives. For Turkish Dutch, however, a different pattern holds, as is clear in the following fragment from an otherwise Dutch conversation of two Turkish young people A and B, while playing cards (where Q (Queen) is the equivalent of Dutch V (Vrouw)):

(14) A: *V **yok**. Is Q. **Tamam mı?***
V there.is.no. is Q. OK Question.Particle
'There is no V. It is Q. OK?'
B: ***Tamam**.*
OK
'OK.'

Turkish is spoken for just a bit, but only with fixed elements such as *yok* 'there is no', *tamam* 'OK' and *mı* (pronounceable as 'muh') 'question particle', highly frequent, almost fixed, elements. The word order of 'is Q' is Dutch, as is the Dutch copula here. Related to this is that inside the Turkish community in the Netherlands the own language is still very much used, but it is very well imaginable that the frequent, fixed elements become a fixture in the ethnolect of Turkish young people in the Netherlands, in particular when speaking among each other (see also Auer 2003 for similar results from Germany). Amoung young Moroccans the shift has been much more rapid (El-Aissati 2001).

The relation between street language and ethnolect could very well lie in that the pronunciation and the sentence intonation in the ethnolect could be taken over in street language, as well as fixed lexical items from different ethnic groups.

This brings us to the perspective of ethnolects in their actual constitution as a multi-faceted, multidimensional phenomenon. Following the schema in Muysken (in prep.), we may assume four dimensions:

(15) a. The L1 dimension, which would correspond to the local variety of the original language;
b. The L1/L2 dimension, characterized by convergence and crossing (Rampton 1995) between different ethnolectal varieties in street language (see also Hewitt 1986);
c. The dimension of universal principles (UP) such as simplification and omission of unstressed functional elements;
d. The L2 dimension, primarily involving approximation to input from the target.

In different varieties of ethnolect, these different dimensions could play a smaller or larger role. This is schematized in (16):

(16)

Local variety of original language	*Convergence and crossing*
L1	L1/L2
Ethnolect	
UP	L2
Simplification from the target	*Approximation to input*

This perspective would suggest a multi-faceted and variable notion of ethnolects (see also the data from Sweden from Boyd and Fraurud 2005), with four dimensions.

To further illustrate the third dimension, *Universal Principles (UP)*, consider simplification. As Van der Sijs (2005: 186–187) notes, a number of morpho-syntactic features occur in a number of different Dutch ethnolects:

a. article omission
b. no gender distinctions in nouns
c. demonstrative in stead of definite article
d. invariant form of the adjective
e. infinitives in stead of finite verb forms
f. copula omission
g. use of *gaan* 'go' as an auxiliary
h. use of *doen* 'do' and *staan* 'stand' as auxiliaries
i. strong in stead of weak pronouns
j. omission of the reflexive pronoun
k. omission of expletive *het* 'it'
l. omission of 'locative' *er* 'there' or use of strong form *daar*
m. tendency towards canonical SVO order rather than the varying word orders of Dutch
n. different selection of prepositions

With a few exceptions these can be viewed in terms of a strategy of simplification. The simplification perspective can plausibly be linked to pidgin- and creole studies, as has been undertaken by Van der Sijs (2005: 189–194), who compared a number of features of Dutch lexifier creoles postulated by Markey (1982) with Dutch ethnolects. The comparison is presented in Table 3, and includes my own comments.

6. Conclusion

The shift perspective to ethnolects is very appealing because it allows immediate comparison of structural features of different ethnic varieties of the national language, e.g. Chicano English and African American Vernacular English, or Moroccan Dutch and Turkish Dutch. In the same vein, it allows comparisons between ethnic and other varieties, standard and non-standard, of the dominant language in a speech community. It further enables precise quantitative descriptive and analytical statements about what makes the ethnolect different from the standard. As a background assumption, it stresses the essential commonalities between the ethnolects and non-ethnolectal varieties.

The appeal of the multidimensional perspective is that it allows us to contextualize the varieties that have emerged in the process of shift within an overall

Table 3. Comparison of a number of features of Dutch lexifier creoles postulated by Markey (1982) with Dutch ethnolects, based on Van der Sijs (2005: 189–194)

	Feature	Classification ethnolects as in Van der Sijs (2005)	Comments Muysken
1	No gender in nouns	+	Variable underdifferentiation
2	No nominal inflection	+	Variable
3	More regular SVO order	+	Tendencies in some sentence types
4	Invariant pronouns for case and gender	±	Variable
5	No passive verb forms	±	Should be compared with non-immigrant peers
6	No morphologically derived comparative and superlative	–	Should be studied for non-frequent adjectives
7	Tense marked with preverbal particles	–	Tendencies towards over-use auxiliaries
8	Nominal plural through addition of plural third person pronoun	–	–
9	No uninflected forms of the verb	–	Possibly a misunderstanding on the part of Van der Sijs or sources cited by her
10	Same verb for existence and possession	–	Highly specific to some creoles
11	A single, preverbal negation	–	Needs further investigation

account of the multilingual repertoires of speakers of a non-dominant language and of the strategies that they employ to make use of this repertoire. It is the complexity of the verbal repertoire and the alternating reliance on strategies of maintenance and shift, convergence, mixing, and simplification that provides the fuel for the rich texture of the ethnolectal varieties. It also will contribute to the issue to which extent shift ethnolects develop at all. It may be that ethnolects are particularly characteristic of communities with patterns of rapid shift.[3]

The latter point bears repeating: even within a shift perspective, the ethnolect does not correspond to a single variety, but covers a range of styles, as the preliminary findings in our own research show, and as becomes clear from e.g. the work of Charry (1983), who discovered quite a while ago already that even ethnolectal variables highly indexical of Surinamese ethnic identity such as bilabial

3. Our ongoing research will be able to help determine whether there is a recognizable Turkish ethnolect in addition to a Moroccan ethnolect.

[w] in Dutch are subject to stylistic variation. Thus even in the shift perspective, the ethnolect must be viewed as strategic, as a stylistic option, rather than as a monolithic whole.

Viewed in this light the differences between the two perspectives becomes primarily: from how far away is the object of research studied. The shift perspective is more microscopic, allowing for detailed phonetic measures, while the multi-dimensional perspective is more macroscopic. Hence the two perspectives are complementary rather than exclusive.

Abbreviations in glosses

AC = accusative case; INF = infinitive; Q = question particle; R = locative case.

References

Aarssen, J. & Backus, A. 2001. Turks. In *Babylon aan de Noordzee. Nieuwe Talen in Nederland*, G. Extra. & J. J. de Ruiter (eds.), 306–328. Amsterdam: Bulaaq.

Androutsopoulos, J. 2001. From the streets to the screens and back again: On the mediated diffusion of ethnolectal patterns in contemporary German. Universität Gesamthochschule Essen: LAUD: Series A, 522. Paper presented at the ICLaVE I Conference, Barcelona June 29, 2000.

Auer, P. 2003. 'Türkenslang'; Ein jugendsprachliger Ethnolekt des Deutschen und seine Transformationen. In *Spracherwerb und Lebensalter*, A. Häcki Buhofer (ed.), 225–264. Tübingen: Francke.

Backus, A. 1992. *Patterns of Language Mixing. A Study of Turkish-Dutch Bilingualism*. Wiesbaden: Otto Harrassowitz.

Backus, A. 1996. Two in One. Bilingual Speech of Turkish Immigrants in the Netherlands. PhD dissertation, Katholieke Universiteit Brabant, Tilburg (Studies in Multilingualism 1. Tilburg: Tilburg University Press).

Bernstein, B. 1966. Elaborated and restricted codes: An outline. *Sociological Inquiry* 36: 254–261.

Bolle, J. 1994. Sranan Tongo – Nederlandse Code-wisseling en Ontlening. MA thesis, University of Amsterdam.

Boumans, L. 1999. The Syntax of Codeswitching: Analysing Moroccan Arabic/Dutch Conversation. PhD dissertation, Radboud University Nijmegen (distributed by Tilburg University Press).

Boumans, L. 2006. The attributive possessive in Moroccan Arabic speech by young bilinguals in the Netherlands and their peers in Morocco. *Bilingualism. Language and Cognition* 9: 213–231.

Boyd, S. & Fraurud, K. 2005. Who is a native speaker? The diversity of language profiles of young people in multilingual urban contexts in Sweden. Paper Presented at ICLaVE 3, Amsterdam June 2005.

Carlock, E. 1978. Prosodic analysis of two varieties of Buffalo English. *The LACUS Forum* 5: 377–382.
Charry, E. 1983. Een sociolinguïstische verkenning van Surinaams-Nederlands. In *De Talen van Suriname*, E. Charry, G. Koefoed & P. Muysken (eds), 138–161. Muiderberg: Coutinho.
Clyne, M. 1967. *Transference and Triggering*. The Hague: Nijhoff.
Clyne, M. 2000. Lingua franca and ethnolects in Europe and beyond. *Sociolinguistica* 14: 83–89.
Cornips, L. 2000. The use of *gaan* + infinitive in narratives of older bilingual children of Morrocan and Turkish descent. *Linguistics in the Netherlands* 17: 57–67.
Craats, I. van de. 2000. Conservation in the Acquisition of Possessive Constructions. PhD dissertation, Tilburg University.
Danesi, M. 1984. Canadian Italian: A case in point of how language adapts to environment. *Polyphony* 7: 110–113.
Danesi, M. 1985. A glossary of lectal terms for the description of language variation. *Language problems and language planning* 89: 115–124.
Doğruoz, S. 2007. Differences in Spoken Turkish in European Immigration Contexts and in Turkey. PhD dissertation, Tilburg University.
Dubois, S. & Horvath, B. M. 2002. Sounding Cajun: The rhetorical use of dialect in speech and writing. *American Speech* 77: 264–287.
El-Aissati, A. 2001. Verandering en verankering: Taal en identiteit bij jonge Marokkanen. In *Een buurt in Beweging: Talen en Culturen in het Utrechtse Lombok en Transvaal*, H. Bennis, G. Extra, P. Muysken & J. M. Nortier (eds), 251–262. Amsterdam: Aksant.
Fought, C. 2003. *Chicano English in Context*. New York NY: Palgrave Macmillan.
Fought, C. 2006. *Language and Ethnicity*. Cambridge: CUP.
Haugen, E. 1956. *The Norwegian Language in America*, 2 Vols. Philadelphia PA: University of Pennsylvania Press.
Hewitt, R. 1986. *White Talk Black Talk: Inter-racial Friendship and Communication amongst Adolescents*. Cambridge: CUP.
Huwaë. R. 1992. Tweetaligheid in Wierden: Het taalgebruik uit een Molukse gemeenschap. MA thesis, University of Amsterdam.
Johanson, L. 1992. *Strukturelle Faktoren in türkischen Sprachkontakten* [Sitzungsberichte der Wissenschaftlichen Gesellschaft an der J. W. Goethe-Universität Frankfurt am Main 29:5]. Stuttgart: Steiner.
Kishna, S. 1979. Lexicale Interferentie in het Sarnami. MA thesis, University of Amsterdam.
Kotsinas, U.-B. 1998. Language contact in Rinkeby, an immigrant suburb. In *Jugensprache, Langue des jeunes, Youth language*, J. K. Androutsopoulos & A. Scholz (eds.), 125–148. Frankfurt: Peter Lang.
Markey, T. L. 1982. Afrikaans: creole or non-creole? *Zeitschrift für Dialektologie und Linguistik* 49: 169–207.
Matras, Y. 2000. Fusion and the cognitive basis of bilingual discourse markers. *International Journal of Bilingualism* 4: 505–528.
Muysken, P., Kook, H. & Vedder, P. 1996. Papiamento/Dutch code-switching in bilingual parent-child reading. *Applied Psycholinguistics* 17: 485–505.
Muysken, P. In prep. Modeling language contact. Bilingualism optimization strategies. Ms, Radboud University Nijmegen.
Nortier, J. 1990. *Dutch-Moroccan Code Switching among Maroccans in the Netherlands*. Dordrecht: Foris.

Nortier, J. 2001. *Murks en Straattaal. Vriendschap en Taalgebruik onder Jongeren*. Amsterdam: Prometheus.
Rampton, B. 1995. *Crossing: Language and Ethnicity among Adolescents*. London: Longman.
Schaufeli, A. 1991. Turkish in an Immigrant Setting. PhD dissertation, University of Amsterdam.
Selting, M. 2006. Einheitenkonstruktion im Türkendeutschen: Grammatische und prosodische Aspekte. *Zeitschrift für Sprachwissenschaft* 25: 239–272.
Selting, M. 2007. Prosody and unit construction in an ethnic style: The case of Turkish German and its use and function in conversation. Panel on Prosody and Pragmatics in Spoken Language Corpora. 10th International Pragmatics Conference, Gothenburg, 8–13 July 2007.
Starren, M. 2001. The Second Time. The Acquisition of Temporality in Dutch and French as a Second Language. PhD dissertation, Tilburg University.
Sijs, N. van der. 2005. *Wereldnederlands. Oude en Jonge Variëteiten van het Nederlands*. The Hague: Sdu.
Thomason, S. & Kaufman, T. 1988. *Language Contact, Creolization and Genetic Linguistics*. Berkeley CA: University of California Press.
Verschik, A. 2007. Jewish Russian and the field of ethnolect study. *Language in Society* 36: 213–232.
Wölck, W. 1984. Sounds of a city: Types and characteristics of the speech of Buffalo and its ethnic groups. *New York Folklore* 10: 7–22.

Applying language technology to detect shift effects

John Nerbonne, Timo Lauttamus, Wybo Wiersma,
Lisa Lena Opas-Hänninen
University of Groningen / University of Oulu / University of Groningen /
University of Oulu

We discuss an application of a technique from language technology to tag a corpus automatically and to detect syntactic differences between two varieties of Finnish Australian English, one spoken by the first generation and the other by the second generation. The technique utilizes frequency profiles of trigrams of part-of-speech categories as indicators of syntactic distance between the varieties. We then examine potential shift effects in language contact. The results show that we can attribute some interlanguage features in the first generation to Finnish substratum transfer. However, there are other features ascribable to more universal properties of the language faculty or to "vernacular" primitives. We also conclude that language technology provides other techniques for measuring or detecting linguistic phenomena more generally.

1. Introduction

The present paper[1] applies techniques from language technology, i.e. application-oriented computational linguistics, to detect syntactic differences between two different varieties of English, those spoken by first and second generation Finnish Australians. It also examines the degree to which the syntax of the first generation differs from that of the second, presumably due to the language shift that the first generation group made later in life and the traces it has left in their English. This line of research naturally attempts not only to detect differences of various kinds, but also to interpret their likely sources, including first language interference but also more general tendencies, called "vernacular primitives" by Chambers (2003: 265–266). To explain differential usage by the two groups, we also draw on

1. This project is partly funded by the Academy of Finland (project # 113501).

the strategies, processes and developmental patterns that second language learners usually evince in their interlanguage regardless of their mother tongue (Færch & Kasper 1983; Larsen-Freeman & Long 1991; Ellis 1994). To forestall a potential misunderstanding, we note that we propose how to automate the **detection** of the concrete syntactic differences, but not their **interpretation** (possible causes). The paper will summarize the findings concerning the Finnish emigrants that Lauttamus, Nerbonnne & Wiersma (2007) report on at length in order to give the reader a sense of the potential of the technique.

A second purpose of the paper is to reflect and generalize on the success of this technique borrowed from language technology in order to suggest that language technology might be a promising source in which to seek techniques for measuring or detecting linguistic phenomena more generally. Language technology has developed a number of techniques which expose the latent structure in language use. We harness one of those in the study of language contact, namely tagging words with their syntactic categories (parts-of-speech, hence POS), in an effort to detect the syntactic differences in the speech of juvenile vs. adult emigrants from Finland to Australia. We shall note other promising opportunities, but the purpose of the reflection is naturally not to claim that the language technology is a panacea for problems of linguistic analysis, but rather to stimulate readers to look toward language technology to explore issues in contact linguistics.

The first part of the paper summarizes the work on detecting syntactic interference among the Finnish emigrants to Australia, and the second makes the more programmatic argument that language technology should not be regarded as a set of tools for applications, but rather as a set of generic tools for exposing linguistic structure. Our paper does not focus on language contact exclusively as this has been influenced by globalization, but the contact effects we focus on do result from a substantial migration from one side of the earth to another. Our intention is to contribute to general techniques for the detection of syntactic differences.

2. Detecting syntactic differences: Techniques

Syntactic theory uses analysis trees showing constituent structure and dependency structure to represent syntactic structure, so a natural tool to consider for the task of detecting syntactic differences would be a parser – a program which assigns the syntactic structure appropriate for an input sentence (given a specific grammar). We decided, however, against the use of a parser, and for the more primitive technique of part-of-speech tagging (explained below) because, even though automatic parsing is already producing fair results for the edited prose of newspapers, we suspected that it would be likely to function very poorly on

the conversational transcripts of second language learners. Both the conversation style of the transcripts and the frequent errors of learners would be obstacles. We return below to the selection of corpora and its motivation.

2.1 Tagging

We can detect syntactic differences in two corpora in a fairly simple way (Lauttamus et al. 2007). We first tag the two corpora automatically, i.e. we detect for each word its syntactic category, or, as it is commonly referred to, its part-of-speech (POS). Below we provide an example:

(1)	*the*	*cat*	*is*
	ART (def)	N (com, sing)	V (cop, pres)
	on	*the*	*mat*
	PREP (ge)	ART (def)	N (com, sing)

We tagged the corpora using the set of POS tags developed for the TOSCA-ICE, which consists of 270 POS tags (Garside et al. 1997), of which 75 were never instantiated in our material. Since we aim to contribute to the study of language contact and second language learning, we chose a linguistically sensitive set, that is, a large set designed by linguists, not computer scientists. In a sample of 1,000 words we found that the tagger was correct for 87% of words, 74% of the bigrams (a sequence of two words), and 65% of the trigrams (a sequence of three words). The accuracy is poor compared to newspaper texts, but we are dealing with conversation, including the conversation of non-natives. Since parsing is substantially less accurate than POS tagging, we feel that this accuracy level confirms the wisdom of not trying to use the more informative technique of full parsing.

The POS tags are then collected into ordered triples, the *trigrams* ART(def)-N(com, sing)-V(cop, pres), ..., PREP(ge)-ART(def)-N(com, sing). We use POS trigrams, rather than single tags, as indications of syntactic structure in order to obtain a fuller reflection of the complete syntactic structure, much of which is determined once the syntactic categories of words are known. In making this last assumption, we follow most syntactic theory, which postulates that hierarchical structure is (mostly) predictable given the knowledge of lexical categories, in particular given the lexical 'head'. Sells (1982, sec. 2.2, 5.3, 4.1) shows how this assumption was common to theories in the 1980s, and it is still recognized as useful (if imperfect given the autonomy of "constructions", which Fillmore & Kay 1999, demonstrate). So if syntactic heads have a privileged status in determining a "projection" of syntactic structure, then we will detect syntactic differences in two varieties by quantifying the distribution of parts-of-speech in context.

2.2 Comparison

We then collected all the POS trigrams found in the corpora (13,784 different POS trigrams in the case of the Finnish Australian data), and counted how frequently each occurred in both of the corpora. We then compared this 2 X 13,784 element table, asking two questions. First, we wished to know whether the distribution in the two rows might be expected by chance, in other words, whether there was a statistically significant difference in the distributions. Second, in case the overall distributions differed significantly (p-values at or below 0.05), we calculated which frequent POS trigrams were responsible for the skewed distribution. We suppress the technical details in this presentation, referring the interested reader to Nerbonne & Wiersma (2006).

In connection with the second goal, we examined the 200 POS trigrams that contribute the most to the skewing of the distribution between the two corpora. Both the relative differences in corpora (i.e. which percentage of a given POS trigram occurs in one corpus as opposed to another) and also the overall frequencies of the trigram were taken into account. We weighted more frequent POS trigrams more heavily because more frequent patterns are likely to be perceptively salient, and also because we are most certain of them. We turn to an examination of the Finnish Australian data below.

2.3 Discussion

By analyzing differences in the frequencies of POS trigrams, we importantly identify not only deviant syntactic uses ("errors"), but also the overuse and underuse of linguistic structures, whose importance is emphasized by researchers on second language acquisition (Coseriu 1970; Ellis 1994: 304–306 uses for underuse 'underrepresentation' and overuse 'over-indulgence'; de Bot et al. 2005: A3, B3). According to these studies, it is misleading to consider only errors, as second language learners likewise tend to overuse certain possibilities and tend to avoid (and therefore underuse) others. For example, de Bot et al. (2005) suggest that non-transparent constructions (such as existential *there* and anaphoric *it* for native speakers of Finnish) are systematically avoided even by very good second language learners.

We like to emphasize that our work does not assume that syntax consists solely of part-of-speech sequences, but only that differences in part-of-speech sequences are indicative of syntactic differences in general. It is important to emphasize that we do not claim to have developed a technique that probes all conceivable syntactic differences directly, but rather a technique that detects traces of differ-

ences in surface syntax. Those differences might naturally have causes in deeper levels of syntactic structure. In a contribution with more room for reflection, we would expand on how we derive inspiration from other indirect measurement techniques such as the measurement of latitude via differences in the local solar noon with respect to Greenwich mean time, or the measurement of temperature via the expansion of a fluid.

Uriel Weinreich (1953: 63) noted the difficulty of aggregating over language contact effects:

> No easy way of measuring or characterizing the total impact of one language on another in the speech of bilinguals has been, or probably can be devised. The only possible procedure is to describe the various forms of interference and to tabulate their frequency.

Our proposed technique for detecting syntactic differences does indeed aggregate over many indicators of syntactic difference, in a way that makes progress toward assessing the "total impact" in Weinreich's sense, albeit with respect to a single linguistic level, namely syntax. We do not develop a true measure of syntactic difference here as that would require further calibration and validation, preferably cross-linguistically, but we do claim to detect differences in the frequency with which different constructions are used.

If such a measure could be validated and calibrated, it would be important not only in the study of language contact but also in the study of second language acquisition. We might then look afresh at issues such as the time course of second language acquisition, the relative importance of factors influencing the degree of difference such as the mother tongue of the speakers, other languages they know, the length and time of their experience in the second language, the role of formal instruction, etc. It would make the data of such studies amenable to the more powerful statistical analysis reserved for numerical data.

2.4 Previous work

Aarts & Granger (1998) suggest focusing on tag sequences in learner corpora, just as we do. We add to their suggestion statistical analysis using permutation statistics, which allows us to test whether two varieties vary significantly. We discuss similar technical work below, none of which has focused on analyzing language contact, however.

3. The Australian English of Finnish emigrants

We shall describe the differences between the English of those who emigrated as adults and those who emigrated as children (juveniles). After studying the transcripts, we conclude that the latter's English is near native, and so we focus below on the English of those who emigrated as adults.

3.1 Linguistic situation of the adult emigrants

We note that the linguistic development of the two Finnish groups in Australia is best described as language shift. We are therefore concerned with bi-generational bilingualism as a series of stages in the assimilation of the Finnish ethnic minorities into a linguistically, socially and culturally English-dominant speech community, which inevitably entails Anglicization among these ethnic groups. Various studies show that language shift usually takes place no later than during the 2nd generation of various ethnic groups in the US, with the exceptions of Spanish and Navajo (Karttunen 1977; Veltman 1983; Smits 1996; Klintborg 1999). The evidence from Hirvonen (2001) also supports this; American Finnish does not seem to survive as a viable means of communication beyond the second generation.

The situation is similar in Australia. Clyne & Kipp (2006: 18) note that "high-shift" groups in Australia tend to be ones who are culturally closer to Anglo-Australians in contrast with some "low-shift" groups with different "core values such as religion, historical consciousness, and family cohesion". The evidence in Lauttamus et al. (2007) suggests that Finnish Australians also form a high-shift group, as they shift to English very rapidly in the second generation. Very few members of the 1st generation of immigrants avoid Finnish interference, (cf. Piller 2002), but members of the 2nd generation usually speak natively (cf. Waas 1996; Schmid 2002, 2004; Cook 2003; Jarvis 2003). Consequently, we expect to find most of the evidence for syntactic interference (substratum transfer) in the English of first generation Finnish Australians, as the second generation has already shifted to English without any interference from Finnish. The findings in Lauttamus et al. (2007) suggest that second generation Finnish Australians speak (almost) natively, with very little Finnish interference in their English. This is corroborated by findings in some other studies, such as Lahti (1999) and Kemppainen (2000) on lexical features, Mannila (1999) on segmental features, Laakkonen (2000) on rhythm, and Markos (2004) on hesitation phenomena.

Like similar groups in the United States (cf. Lauttamus & Hirvonen 1995), the adult immigrants typically go on speaking Finnish at home as long as they live, and carry on most of their social lives in that language, leaving Finnish their

dominant language. They struggle to learn English, with varying success, e.g. usually retaining a noticeable foreign accent. But they are marginally bilingual, as most of them can communicate successfully in English in some situations.

We contrast their situation with that of their children. The immigrant parents speak their native language to their children, so this generation usually learns the ethnic tongue as their first language. The oldest child may not learn any English until school, but the younger children often learn English earlier, from older siblings and friends. During their teens the children become more or less fluent bilinguals. Their bilingualism is usually English-dominant: they tend to speak English to each other, and it is sometimes difficult to detect any foreign features at all in their English. As they grow older and move out of the Finnish communities, their immigrant language starts to deteriorate from lack of regular reinforcement. Even if this generation marries within its ethnic group, as is frequently the case, English nonetheless becomes the language of the household, and only English is spoken to the following generation.

Language contact scholarship distinguishes situations of *shift* from *maintenance* (Thomason and Kaufmann, 1988; Van Coetsem, 1988). The adult emigrant group, our focus here, shifts to English. Their Finnish is linguistically dominant, while English is socially dominant throughout Australia. In a situation of adult language shift, we expect interference from the native (Finnish) in the acquired (English) language, beginning with pronunciation (phonology) and morphosyntax. Lexical interference is comparatively weak.

3.2 Finnish Australian English Corpus (FAEC)

Greg Watson of the University of Joensuu compiled a corpus of English conversations with Finns who had emigrated to Australia nearly thirty years earlier (Watson 1996). This corpus was kindly put at our disposal. All the respondents were Finnish native speakers. We divided them into two groups, "adults", or adult emigrants, who were over 18 upon arrival in Australia, and "juveniles", the children of the adults, who were all under 17 at the time of emigration. We distinguish between adult immigrants and immigrant children based on Lenneberg's (1967) well-known critical age hypothesis, which suggests a possible biological explanation for successful L2 acquisition between age two and puberty. Note that 'adult' vs. 'juvenile' refers only to the age at emigration: all the respondents were over 30 at the time of the interviews.

The adults' average age was 30 years on arrival, while the children's was six, but the older group was 58.5 at the time of the one-hour interview, and the younger one 36. There were 62 adults and 28 juveniles interviewed, and there were roughly

equal numbers of males and females. The interviews were transcribed in regular orthography by trained language students and later checked by Watson. Speakers were not tested for English proficiency, but it is clear from a quick view of the data that the juveniles' English is considerably better than that of the adults'. The juveniles had gone to school in Australia, and the adults in Finland. Our corpora contained 305,000 words in total.

4. Differences observed

The following section summarizes some of the material in Lauttamus et al. (2007). The evidence from our syntactic analysis using the POS-tag trigrams and a permutation test like the one described in detail in Nerbonne & Wiersma (2006) shows that there are differences between the adults and the juveniles at a statistical significance level of 0.01. Our report focuses first on the aggregate effects of syntactic distance between the two groups of speakers and then we move on to discuss more specific "syntactic contaminants" in the English of the adults. The role of the language technology, specifically the POS-tags and the permutation test used to identify differing elements in the distribution, is that of detection. We also attempt to interpret the differences, but we have not enlisted language technology for this purpose.

4.1 General effects

Some of the significant syntactic differences found in the data might be attributed to the lower level of fluency in the adults. Their language exhibits the following:

a. Overuse of *hesitation phenomena* (pauses, filled pauses, repeats, false starts etc.), arising from difficulties in speech processing and lexical access.
b. Overuse of *parataxis* (particularly with ***and*** and ***but***) as opposed to hypotaxis.
c. Underuse of *contracted forms* that the juveniles use easily and naturally, e.g. *I've been running, I'd like to go, I'll finish my degree.* Adults mostly use full forms such as *I have been, she will be.*
d. Reduced repertoire of *discourse markers* such as *you know, you see, I mean.* Adults do use *you know* (with other hesitation phenomena), but as a time-gaining device rather than as a genuine discourse marker. In contrast, the juvenile emigrants use a more varied repertoire of markers, which often function as appeals to the interviewer.

e. Avoidance of *complex verb clusters.* Juvenile, but not adult, emigrants use structures such as *I would have had it, I still probably would have ended up getting married.*
f. Avoidance of *prepositional and phrasal verbs.* In contrast, juvenile emigrants have no difficulty with verbs such as *I* ***ran out of*** *money, I just* ***opted out for*** *an operation.*
g. Underuse of the existential *there.* The adults either avoid using the existential or attempt to express it without the word *there* (cf. Section 4.2.3). We include this in the list of general differences as an example of a general difficulty that speakers have with peculiar (non-transparent) English constructions.

We extracted the properties above by investigating frequent POS-trigrams that differed significantly in one group as opposed to another (individual *p*-values at or below 0.05), also using 90% relative frequency as a threshold, i.e. where 90% occurred in the one group or the other (once we applied a correction for overall difference in corpus size). This means that avoidance does not imply total absence of a feature in a group. Nor do we wish to suggest that adults are consciously avoiding certain constructions such as hypotaxis. The differences in usage patterns could arise through other strategies.

The ability to identify these sources of deviation in the use of English by the adult Finnish emigrants confirms our contention that the comparison of POS-trigram frequencies indeed reflects the syntactic distance between the two varieties of English and, consequently, aggregate effects of the difference in the two groups' English proficiency. The shift to English has indisputably proceeded along different paths in the two groups, the adults (still) showing features of "learner" language, or shift with interference, and the juveniles those of shift without interference.

4.2 Specific syntactic effects

We turn to differences in specific constructions. In examining these, we shall interpret them on the basis of our knowledge of standard (acrolectal) Finnish and English, which is a risky undertaking. We shall likewise entertain interpretations based on what we know about non-standard (basilectal) varieties of English and Finnish, but our knowledge is less than perfect here.

In examining the following differences in POS-trigram frequency, we will be asking whether the observed syntactic deviations from the norms of standard (acrolectal) English may be ascribed to contact effects from a Finnish substratum, to more universal, 'natural' tendencies in non-standard varieties in general, or to other factors. Modesty compels us to note that we are aware that there are many further sources of influence which might explain why the language of second

language learners differs from that of native speakers. Lauttamus et al. (2007) discusses this in more detail.

To illustrate how explanations compete, consider the fact that different adult speakers fail to enforce subject-verb concord, thus *They all doned here, they, – they wasn't raw [kangaroo] skin.* The subject-verb nonconcord in *they wasn't* ('they weren't') is putatively a vernacular universal. But in non-acrolectal Finnish, subject-verb nonconcord is also frequent, e.g. *ne meni Groningeniin* ('they went to Groningen', *ne*, plural of *se* 'it', + *meni* 'went' 3rd person SG), which shows subject-verb nonconcord in person and number, as opposed to standard Finnish: *he menivät Groningeniin* (*he* 'they' + *meni+vät* 'went' + 3rd person PL). Just as some vernacular Englishes, non-acrolectal Finnish also violates the standard subject-verb concord rule.

To support a potential role of the 'vernacular' approach in our analyses, we refer to Fenyvesi & Zsigri (2006: 143). They suggest that less educated speakers of English (such as the adults), who have usually learnt their L2 via listening, rely on *auditory* input, whereas more educated immigrant language speakers (such as our juveniles), who have acquired their L2 also through reading and writing, and therefore been exposed to a more or less codified standard (acrolectal) variety, rely on *visual* input as well. The fact that the adults in our study have mainly been exposed to spoken, *basilectal* (Australian) English is likely to give rise to some general vernacular features.

We discuss two patterns in detail, one we attribute to the Finnish substrate, and the other to general simplifying tendencies. We note several others more briefly, hoping to provide the flavor of the previous work.

4.2.1 *Article usage*

The adults demonstrate overuse (and underuse) of the indefinite and definite articles, *a(n)* and *the*, characteristic of a learner whose L1 has no article system (such as Finnish), as exemplified in the following:

(2a) *in that time /in a Finland/ because wasn't very*
(2b) *first we go /to the Finland/*
(3) *we been /in a Brisbane/ Brisbane because ah*
(4) *in /the Brisbane and/*
(5) *I had /a different birds/ in Finland*

In example (5) the indefinite article occurs with a countable, plural noun head, a very unlikely overuse for a native speaker, however informal. The juveniles do not show similar linguistic behavior being more proficient in English.

Finnish Americans also overuse the articles, particularly the indefinite article, using it, for example, with proper nouns. Pietilä (1989: 167–168) shows that Finnish Americans, particularly elderly, first generation speakers often supply

redundant definite articles, such as those in (2b) and (4) rather than indefinites, such as (2a) or (3). Pavlenko & Jarvis (2002: 207) show that most of the L1-influenced article errors committed by Russian L2 users of English were omissions and that only a few involved oversuppliance of the definite article. (Similarly to Finnish, Russian has no article system.)

Finnish does allow for the use of the demonstrative pronouns *this* and *that* to mark definiteness instead of the definite article, which may explain the overuses we found in English demonstratives used by the adult emigrants:

(6) *it's /this taxation is/ really something in Finland*
(7) *I watch /that ah news/ and 'Current Affair'*

In these contexts there is no apparent need to use the demonstratives, e.g. a need to contrast one news (broadcast) with another. We note, however, that in a potential Finnish variant of (6), *juuri tämä verotus* […] *Suomessa*…('[it's] the very taxation […] in Finland') it would be quite acceptable to use the demonstrative *tämä* 'this' to make the reference not only definite but also specific. We conclude tentatively that the adult overuse of the demonstratives also originates from Finnish substratum transfer. This is consonant with the fact that adults may also overuse *that one* in expressions such as *I don't /remember that one/ either, I can't /explain that one/, I can't really /compare that one/*, where the NP *that one* has more or less the same function as the pronoun *it*.

To summarize, we ascribe the deviant usage of the articles in the English of the adult Finnish Australians to substratum transfer from Finnish (which has no article system to express (in)definiteness and specificity). Because Finnish has no articles, we might think that there is nothing to transfer (cf. Arabski 1979). However, we agree with Ellis (1994: 306–315), who argues that the absence of a feature in the first language may have as much influence on the second language as the presence of a different feature. In addition to contact-induced effects, it appears that general *hypercorrection* (or overgeneralization), common in 'learner' language, may be a contributing factor. In this light, an uncertainty of article usage in speakers whose L1 has no articles is "universal".

4.2.2 *Acquired formulae*

The distribution of the POS-trigrams also revealed that the adults have acquired some formulae such as *that's* and *what's* without mastering their grammar:

(8) *ah* /that's is/ *not my occupation*
(9) *I think* /that's is/ *a no good*
(10) *um* /that's is/ *a same um*
(11) *and* /that's a/ *causing discomfort in*
(12) *oh* /what's is/ *on that*

That's and *what's*, acquired as fixed phrases, have apparently been processed as single elements. The fact that they are then combined with full copulas or progressive auxiliaries indicates that the speaker has not mastered the grammar of the reduced-form clitic. We also found examples such as *what's a that sign, what's a that seven or something.* Ellis (1994: 20), for one, argues that learners often produce formulae or ready-made chunks as their initial utterances. Acquired formulae cannot be ascribed to substratum transfer, as they tend to be recurrent in any interlanguage. In particular, however, we know of no plausible Finnish model for this difference.

4.2.3 *Other deviant patterns detected via POS-tag distribution differences*

In this section we note some of the other deviant patterns discussed at length in Lauttamus et al. (2007). We discuss them here to add a sense of the value of the technique.

Omission of the progressive auxiliary ***be****.* Adults frequently omit the progressive auxiliary verb *be* (present and past) while the juveniles do not. The adults produce examples such as *when we /drivin' in the/ road.* Absence of the copula *be* is one of the alleged vernacular universals (Chambers 2004), and we also found more numerous examples of that in the adults' speech. Even though Finnish has no formal contrast between the progressive and non-progressive aspect, so that it might be expected to be problematic, still that does not seem to explain this specific error, which we therefore would ascribe to more universal properties of language contact rather than to substratum transfer from Finnish, specifically the difficulty learners have in acquiring unstressed elements. Pietilä (1989: 180–181) also notes that the most frequent verb form error in the English of the first generation Finnish Americans is the omission of the primary *be* in the progressive.

Omission of existential ***there*** *and anaphoric* ***it***. We noted the possible omission of existential (expletive) *there* in 4.1 above, which we class here with the deviant use of anaphoric *it* in subject position. We find examples such as *and summer /time when Ø is/ a people*, where it is apparent that the speaker is aiming at *when* ***there*** *is/are people* or *because is [tough]* where '*because* ***it*** *is [tough]*'. These examples can be explained in terms of substratum influence from Finnish, which would assign the subject argument of the copula verb *be* to the NP *(a) people*, or to the AP *[tough]*, and, consequently, would not mark the subject in the position before the copula.

Absence of prepositions. The adults tend to leave out prepositions with motion verbs such as *move*, *go*, *come*, as exemplified in *and they move me /****other room*** *where/*. Since Finnish has no prepositions, this looks like a straightforward case of

substratum influence, but the case is complicated by the fact that many vernacular varieties of English tend to leave out prepositions in expressing spatial relations with motion verbs (cf. e.g. Linn 1988).

Deviant word order. The adults also demonstrate deviant word order, particularly with adverbials, which are often placed before the object, as exemplified in *I / don't like really/ any old age*. As pre-object placement of the equivalent adverbials in Finnish would be quite acceptable, we suspect that this is contact-induced. Similarly time adverbials were found to be positioned differently, with the adults favoring a front position for frame adverbials such as *fifteen years ago we drivin' around*, but not so drastically as to result in out-and-out errors. This is a feature that can be ascribed to Finnish substratum transfer as well, since in Finnish a time adverbial often appears in this position, without being accompanied by focus as in English. We conjecture that the adults are overusing this construction ignoring its pragmatic conditioning.

***Not** in pre-verbal position*. The adults produce negated utterances where the *not* is placed in pre-verbal position, as in *but uh /we not cook/ that way* (without the supporting verb *do*). This can be ascribed to Finnish substratum transfer, because Finnish always has the negative item (*ei*, inflected in person and number like any verb in standard Finnish) in pre-verbal position. Here again, there is a plausible alternative explanation from universal tendencies in second language syntax. Ellis (1994: 99–101, 421–422) notes that "there is strong evidence that in the early stages of L2 acquisition learners opt for preverbal negation, even where the L1 manifests postverbal negation" (p. 421).

4.2.4 *Conclusions with respect to results*

Lauttamus et al. (2007) note as well that differences became visible with respect to the use of *what* as a relative pronoun (*cars what they built*), the overuse of the simple present (*when we come in Australia thirty years ago*), and in the tendency to form all measure terms with plurals; we found only one example in which the singular is used (*three **foot** wide standing up*). The conclusion there was that the computational techniques had been useful in detecting deviant patterns, but the paper was focused on the interpretation of differences found.

We note that our work has focused on detecting differences using rough, shallow syntax annotation. A calibrated metric of difference would provide a numeric score of syntactic difference in a way that would allow us to compare across languages, e.g. to compare the English of Finnish immigrants to the English of Chinese immigrants, and we have not attempted that to date. It is clear that this would be interesting from several points of view, including second language learning

and language contact studies. One might think that the R^2 scores (or χ^2 scores) used internally might serve this purpose, but they both depend on corpus size, which is a poor property for a candidate measure. Minimally we would need to correct for corpus size in order to use them. This will be future work.

5. Language technology offers tools to study language contact

There are several similar uses of language technology (LT) in detecting differences in language use, especially in information retrieval, authorship detection, and forensic linguistics, in general in all those fields where *text classification* plays a role. Information retrieval classifies texts into relevant vs. irrelevant (with respect to a user query), authorship detection classifies texts according to their authors, and forensic linguistics does the same (among other things). Nerbonne (2007) reviews especially the work on authorship, surely the most challenging classification task. The suggestion here is that we might approach language contact studies from a similar perspective, attempting to design systems that classify texts into native and contact-affected, naturally always from the perspective of a single language (we are unlikely to detect contact effects cross-linguistically). It almost goes without saying that for language contact studies the real interest is less in the sheer ability to classify and more in the linguistic features that form the basis of the classification, but the other fields have likewise been interested in the linguistic basis of successful classifiers, so language-contact studies is by no means alone in the wish to identify concrete linguistic effects. The well-informed reader may object that a great deal of text classification focuses on lexical features, but we would note that syntactic features enjoy growing popularity (see below as well).

As a brief aside, let us add that there has been a significant infusion of LT techniques in the field of dialectology, which (often) attempts to separate varieties into classes of dialects spoken in dialect areas, and which has therefore made use of classifying techniques similar to those used in text classification. Nerbonne & Heeringa (to appear) provides an overview of LT techniques which have been brought to bear on measuring varietal differences, and Spruit (2008) applies LT techniques to the problem of classifying Dutch dialects on the basis of syntax. Spruit had the luxury of basing his work on an elaborate database on the syntactic properties of the Dutch dialects, the *Syntactic Atlas of Netherlandic Dutch* (SAND). We are not aware of similar resources available in language contact studies.

If the study reported on in the first part of this paper appears promising, we suggest that further investigations into the use of LT in contact studies would be fruitful. To give some idea of what might be possible, let us discuss points at which the present article might have been different, perhaps better.

For example, we might want to detect syntactic structure at a more abstract level. We chose to use POS-tags, the syntactic categories of the words in the text. But LT offers at least three more possibilities, that of *chunking*, that of *partial parsing*, and that of *full parsing*. Chunking attempts to recognize non-recursive syntactic constituents, *[the man] with [the red hair] in [the camel-hair jacket] waved as he [passed by],* while partial parsing attempts to infer more abstract structure where possible: *[$_S$ [$_{NP}$[the man] with [the red hair]] in [the camel-hair jacket] waved] as [$_S$ he [passed by]]*. Note that the latter example is only partially parsed in that the phrase *[in the camel-hair jacket],* which in fact modifies *man,* but which might mistakenly be parsed to modify *[hair],* is not attached in the tree. In general, chunking is easier and therefore more accurate than partial parsing, which, however, is a more complete account of the latent hierarchical structure in the sentence. There is a trade-off between the resolution or discrimination of the technique and its accuracy.

Baayen, van Halteren & Tweedie (1996) work with full parses on an authorship recognition task, while Hirst & Feiguina (2007) apply partial parsing in a similar study, obtaining results that allow them to distinguish a notoriously difficult author pair, the Brontë sisters. The point of citing them here is to emphasize that LT methods are being applied to scholarly problems even today: one should not regard them merely as promising possibilities for the future.

Our study could also have used these more discriminating techniques, but we were dissuaded by the fact that the more sensitive techniques have more difficulty in analyzing unedited, indeed, very spontaneous, text, which has the added difficulty of being riddled with second language errors. But whether this is in fact the case is an empirical question waiting for a future research project.

It is clear that the technical view may add to the value of the work. Hirst & Feiguina (2007) are at pains to establish that their technique can work for even short texts (500 wd. and fewer), and this could be an enormous advantage in considering other applications, e.g. to the pedagogical question of identifying foreign influences in the writing of second language users.

6. Conclusion

In this paper we argue that by using frequency profiles of trigrams of POS categories as indicators of syntactic distance between two different groups of speakers we can detect the "total impact" of L1 on L2 in SLA. Our findings show syntactic contamination from Finnish in the English of the adult first generation speakers, and, moreover, we were able to identify several syntactic areas in which the adult emigrants differed significantly from their native-like children. Some of the

features found in the data can be explained by means of contact-induced influence whereas others may be primarily ascribed to "learner" language or to more universally determined properties of the language faculty. We close the paper with an appeal to researchers in the study of language contact to look to language technology for tools to reveal the latent structures in language use, especially syntactic and phonological structure.

References

Aarts, J. & Granger, S. 1998. Tag sequences in learner corpora. A key to interlanguage grammar and discourse. In *Learner English on Computer*, S. Granger (ed.), 132–141. London: Longman.

Arabski, J. 1979. *Errors as Indicators of the Development of Interlanguage.* Katowice: Universytet Slaski.

Baayen, H., van Halteren, H. & Tweedie, F. 1996. Outside the cave of shadows. Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing* 11(3): 121–131.

Bot de, K., Lowie, W. & Verspoor, M. 2005. *Second Language Acquisition: An Advanced Resource Book.* London: Routledge.

Chambers, J. K. 2003. *Sociolinguistic Theory. Linguistic Variation and its Social Implications.* Oxford: Blackwell.

Chambers, J. K. 2004. Dynamic typology and vernacular universals. *Dialectology meets Typology*, B. Kortmann (ed.), 127–145. Berlin: Mouton.

Clyne, M. & Kipp, S. 2006. Australia's community languages. *International J. Soc. Lang.* 180: 7–21.

Cook, V. (ed.). 2003. *Effects of the Second Language on the First.* Clevedon: Multilingual Matters.

Coseriu, E. 1970. *Probleme der kontrastiven Grammatik.* Düsseldorf : Schwann.

Ellis, R. 1994. *The Study of Second Language Acquisition.* Oxford: OUP.

Fenyvesi, A. & Zsigri, G. 2006. The role of perception in loanword adaptation. The fate of initial unstressed syllables in American Finnish and American Hungarian. *SKY Journal of Linguistics* 19: 131–146.

Fillmore, C. & Kay, P. 1999. Grammatical construction and linguistic generalizations. The *what's x doing y* construction. *Language* 75: 1–33.

Færch, C. & Kasper, G. 1983. *Strategies in Interlanguage Communication*. London: Longman.

Garside, R., Leech, G. & McEmery, T. 1997. *Corpus Annotation. Linguistic Information from Computer Text Corpora.* London: Longman.

Hirst, G. & Feiguina, O. 2007. Bigrams of syntactic labels for authorship discrimination of short texts. *Literary & Linguistic Computing* 22(4): 405–419.

Hirvonen, P. 2001. Doni finished – meni läpi – highskoulun. Borrowing, code-switching and language shift in American Finnish. *Global Eurolinguistics. European languages in North America – Migration, maintenance and death,* P. Sture Ureland (ed.), 297–324. Tübingen: Niemeyer.

Jarvis, S. 2003. Probing the effects of the L2 on the L1. A case study. *Effects of the Second Language on the First*, V. Cook (ed.), 81–102. Clevedon: Multilingual Matters.

Karttunen, F. 1977. Finnish in America. A case study in monogenerational language change. In *Sociocultural Dimensions of Language Change*, B. G. Blount & M. Sanches (eds.), 173–184. New York NY: Academic Press.

Kemppainen, J. 2000. Lexical Features in the Spoken English of Finnish Australians. MA thesis, University of Oulu.

Klintborg, S. 1999. *The Transience of American Swedish.* Lund: Lund University Press.

Laakkonen, K. 2000. A Study of the Realization of Rhythm in the English of Finnish Australians. MA thesis, University of Oulu.

Lahti, H. 1999. *Lexical Errors in the Spoken English of Finnish Australians.* MA thesis, University of Oulu.

Larsen-Freeman, D. & Long, M. H. 1991. *An Introduction to Second Language Acquisition Research.* London: Longman.

Lauttamus, T. & Hirvonen, P. 1995. English interference in the lexis of American Finnish. *The New Courant* 3: 55–65. (Department of English, University of Helsinki: Helsinki University Press).

Lauttamus, T., Nerbonne, J. & Wiersma, W. 2007. Detecting syntactic contamination in emigrants. The English of Finnish Australians. *SKY Journal of Linguistics* 21: 273–307.

Lenneberg, E. 1967. *Biological Foundations of Language.* New York NY: John Wiley.

Linn, M. 1988. The origin and development of the Iron Range Dialect in Northern Minnesota. *Studia Anglica Posnaniensia* XXI: 75–87.

Mannila, T. 1999. A Study of Phonic Interference from Finnish in the English of Finnish Australians. MA thesis, University of Oulu.

Markos, M. 2004. 'No, No Swearing, no Swearing Allowed'. A Comparative Study of Hesitation Phenomena in the Spoken English of Two Generations of Finnish Australians. MA thesis, University of Oulu.

Nerbonne, J. 2007. The exact analysis of text. Foreword to the 3rd edn. In *Inference and Disputed Authorship: The Federalist,* F. Mosteller & D. Wallace, xi–xx. Stanford CA: CSLI.

Nerbonne, J. & Heeringa, W. To appear. Measuring dialect differences. In *Theories and Methods*, J. E. Schmidt & P. Auer (eds.). Berlin: Mouton.

Nerbonne, J. & Wiersma, W. 2006. A measure of aggregate syntactic distance. In *Linguistic Distances*, J. Nerbonne & E. Hinrichs (eds), 82–90. Shroudsburg PA: ACL.

Pavlenko, A. & Jarvis, S. 2002. Bidirectional transfer. *Applied Linguistics* 23: 190–214.

Pietilä, P. 1989. *The English of Finnish Americans with Reference to Social and Psychological Background Factors and with Special Reference to Age* [Turun yliopiston julkaisuja, Sarja B, Osa 188]. Turku: Turun yliopisto.

Piller, I. 2002. Passing for a native speaker. Identity and success in second language learning. *Journal of Sociolinguistics* 6: 179–206.

Sells, P. 1982. *Lectures on Contemporary Syntactic Theories.* Stanford CA: CSLI.

Smits, C. 1996. *Disintegration of Inflection. The Case of Iowa Dutch.* The Hague: HAG.

Schmid, M. 2002. *First Language Attrition, Use and Maintenance. The case of German Jews in Anglophone Countries*. Amsterdam: John Benjamins.

Schmid, M. 2004. Identity and first language attrition. A historical approach. *Estudios de Sociolingüística* 5: 41–58.

Spruit, M. 2008. Quantitative Perspectives on Syntactic Variation in Dutch Dialects. PhD dissertation, University of Amsterdam.

Thomason, S. G. & Kaufman, T. 1988. *Language Contact, Creolization, and Genetic Linguistics*. Berkeley CA: University of California Press.

Van Coetsem, F. 1988. *Loan Phonology and the Two Transfer Types in Language Contact*. Dordrecht: Foris.
Veltman, C. 1983. *Language Shift in the United States.* Berlin: Mouton.
Waas, M. 1996. *Language Attrition Downunder. German Speakers in Australia*. Frankfurt: Peter Lang.
Watson, G. 1996. The Finnish-Australian English corpus. *ICAME Journal: Computers in English Linguistics* 20: 41–70.
Weinreich, U. [1953] 1964. *Languages in Contact*. The Hague: Mouton.

Generational differences in pronominal usage in Spanish reflecting language and dialect contact in a bilingual setting

Ricardo Otheguy, Ana Celia Zentella and David Livert
City University of New York / University of California /
Pennsylvania State University

The alternation between presence and absence of subject personal pronouns in Spanish is studied in the bilingual setting of New York City with data extracted from the Otheguy-Zentella corpus. The speech of newcomers to NYC shows that the Caribbean and the Latin American Mainland resemble each other but that there are statistically significant differences between the two regions in occurrence rates of overt pronouns and in the role played by the different person-numbers of the verb in motivating their use. Clear changes in usage are observed in the second, NY-raised generation, consisting of large increases in occurrence rates and in changes in the role played in the selection of overts by the different person-numbers of the verb.

1. Preliminaries

1.1 Subject personal pronouns in Spanish in New York

The alternation between the presence and absence of subject personal pronouns in finite or tensed verbs in Spanish (for example, *yo canto* ~ *canto* 'I sing', or *ella canta* ~ *canta* 'she sings') has been the topic of numerous dialectological and sociolinguistic studies (Bentivoglio 1987, Cameron 1993, 1995, Enríquez 1984, Silva-Corvalán 1982, 1997a, 1997b). This alternation between what are usually called 'overt' (*yo canto*) and 'null' (*canto*) subject personal pronouns is one of the grammatical features regularly covered in overviews of Spanish dialectology at continental or regional levels (López-Morales 1992: 137ff, Lipski 1994: 241ff). The alternation between overt and null subject personal pronouns has also been studied in some detail in language-contact situations, especially in bilingual

Spanish-English areas of the United States (Bayley & Pease-Alvarez 1997, Lipski 1996, Montrul 2004, Silva-Corvalán 1994, 1998, Toribio 1994, 2000, 2004, Zapata, Sánchez & Toribio 2005). Different approaches have been taken to the study of the variable use of subject pronouns in Spanish, but the more frequent one has been quantitative, and the preferred type of data has been naturalistic speech samples derived from narrative retellings or oral sociolinguistic interviews. A recent example of such data is the Otheguy-Zentella corpus, developed at the Graduate Center of the City University of New York (CUNY) for the purpose of documenting and analyzing the Spanish spoken in the complex bilingual and multidialectal environment of New York City (Lapidus & Otheguy 2005a, 2005b, Otheguy & Zentella 2007, Otheguy, Zentella & Livert 2007, Shin & Otheguy 2009). The corpus contains more than 150 hours of transcribed conversations, from which we have extracted and coded more than 60,000 tensed verb tokens, which have been used for the study of variability in the occurrence of subject pronouns.[1] The interviews that form the corpus constitute a stratified sample of 141 Spanish-speaking residents of New York City (NYC) who encompass a wide variety of linguistic profiles.[2] From the point of view of their regional origins, our interviewees hail from six different countries, encompassing what are traditionally regarded as two main dialectal zones in New World Spanish: the Caribbean and the Latin American Mainland. From the point of view of immigrant generations, our consultants range from those recently arrived in NYC, who have had little or no exposure to English or to varieties of Spanish other than their own, to those born in the City who have spoken both English and Spanish from earliest infancy, and have had life-long exposure not only to English but to the many forms of Spanish spoken in the polyglot North American metropolis.

This paper studies the alternation between overt and null subject personal pronouns in regional and generational sub-samples of the Otheguy-Zentella corpus,

1. For the development and coding of the corpus, the authors gratefully acknowledge the support of the National Science Foundation (Grant No. BCS 0004133), as well as preliminary support from the Professional Staff Congress of the City University of New York (Grant No. 62666-00-31) and from the City University of New York (Collaborative Grant No. 09-91917). The authors wish to thank the members of our enthusiastic research and support team at CUNY: Magda Campillo, Eulalia Canals, Lionel Chan, Itandehui Chávez, Daniel Erker, Nydia Flores-Ferrán, Eva García, Manuel Guerra, Karina Hernández, Naomi Lapidus Shin, Óscar Osorio, Silvia Rivero, Jeannette Toro de la Rosa, Juan Valdez, and Zoe Schutzman. We also wish to thank Erika Troseth and Rachel Varra for their help in locating references and editing our work.

2. The corpus has evolved somewhat since its earliest use in analyses such as those reported in Otheguy & Zentella (2007) and in Otheguy, Zentella & Livert (2007). One interview has been eliminated to improve age stratification, so that there are now 141 instead of 142 interviews. Several omissions regarding verbs used with the pronoun *uno*, *una* have also been corrected.

with an eye to documenting language- and dialect-contact phenomena in the Spanish spoken in NYC.[3] The information provided here is expressed in terms of the frequency of occurrence of overt pronouns which, as we shall see presently, increases significantly in the span of one apparent-time generation in NYC. From among several grammatical variables that probabilistically condition the distribution of overt pronouns, we have singled out for analysis the person-number of the verb, because it is with respect to this variable that we find both the clearest regional differences among those of our consultants raised in Latin America (first-generation) as well as the clearest evidence of regional convergence among those raised in New York (second-generation). We propose explanations for the increase in the use of overt pronouns alongside the convergence with respect to the different person-numbers of the verb in terms of language and dialect contact influences.

1.2 Grammatical-discourse variables: Person of the verb

In this as in other studies of subject personal pronouns in Spanish, several grammatical-discourse variables have been found to probabilistically constrain pronoun variability. These include, in addition to the person-number of the verb, the discourse status of its subject as switched or same in reference with respect to the subject of the immediately preceding verb, its particular type of lexical meaning, and the type of clause where it occurs. These constraining factors are the study's independent variables. (The dependent variable is, of course, the occurrence of an overt or null pronoun.) We use the shorthand terms 'Person,' 'Continuity,' 'Lexical,' and 'Clause' to refer to these independent variables.

In a logistic regression analysis that ranked these independent variables, plus several others, in terms of their relative effect on the dependent variable, Otheguy, Zentella & Livert (2007: 789) found that Person was the most influential. Saying it another way, Person accounts for more of the variance between occurrences of overt and null pronouns in the 60,000+ verb tokens of the corpus than any other independent variable. This simply means that the strongest statistical predictor of whether verb tokens appear with an overt or a null pronoun is whether the verb's inflection is first, second, or third person, and that other statistical predictors, such as whether the verb is same or switched in reference, account for lesser amounts of variance.

3. Most of the present study is based on a subset of the corpus, consisting of 67 consultants, selected for the reasons and according to the criteria outlined below, who together produced nearly 30,000 verbs.

Within the variable Person, we study which of its factors statistically favor the use of overt pronouns and which factors disfavor it (that is, favor the use of null pronouns) and to what extent. In other words, we ask whether tokens of 1SG verbal inflections in *–o* favor or disfavor the use of an overt *yo,* whether 2SG inflections in *–as* and *–es* favor or disfavor the use of an overt *tú*, etc.; and we compare whether the favoring of overts by *–o* inflections is greater or less than the favoring of overts by *–as / –es* inflections, and so forth.

1.3 Sociodemographic variables: Region and Generation

The paper focuses on the two most basic independent social variables affecting the distribution of overt and null pronouns in the multilingual and multidialectal immigrant environment of New York City:

a. The speaker's affiliation (because of his or her own, or his or her parents' place of birth) with either the Spanish-speaking Caribbean or the Spanish-speaking Latin American Mainland. We use the shorthand name 'Region' to refer to this variable.
b. The speaker's status as either newcomer to New York or as born or raised in the City (New York-raised or NYR for short). We use the shorthand name 'Generation' to refer to this variable.

With regard to the variable Region, segmenting out a Caribbean area for separate study is standard procedure in Latin American dialectology (Lipski 1994, López-Morales 1992) and justifies our breaking up our NYC sample into Caribbeans and Mainlanders. In the corpus, Caribbeans are consultants with origins in Cuba, Dominican Republic, or Puerto Rico; Mainlanders are consultants with origins in the non-coastal regions of Colombia, Ecuador, or Mexico. (Thus, the countries that our consultants come from include the largest Spanish-speaking country in the world, Mexico, and the fourth largest, Colombia, as well as the one with the largest representation in New York City, Puerto Rico.) With regard to the variable Generation, newcomers are consultants born in Latin America who came to New York after age 16 and have lived in the City for less than six years; NYR are consultants who were born in New York or were brought to the City at age three or younger.

Our criteria for defining newcomers were designed to provide a subsample of first-generation speakers who have had the least exposure to English and to forms of Spanish other than their own. Our newcomers arrived in New York with a solidly formed native Spanish, have resided for a relatively short time in the City, and missed spending in NYC the highly formative years of teen-age sociability and

high school education. Our criteria for establishing the group of NYR provide just the opposite, that is, a subsample of second-generation speakers who, having been born in New York or having been brought to the City in early infancy, have had the most exposure to English and other varieties of Spanish, and have had their entire education in NYC.

1.4 Separate bivariate analyses

In the analysis of the variable use of subject pronouns among speakers from different Latin American regions and different NYC immigrant generations, this paper relies on serial bivariate analyses, using percentages and cross-tabulations. No attempt is made to represent the grammatical-discourse results in a multivariate approach. While multivariate analyses are standard in sociolinguistics and do indeed offer a truer picture of the distribution of forms, such analyses remain less than transparent for many scholars, who often ask about, and are more likely to make use of, the simpler type of bivariate information provided here. An in-depth multivariate analysis of part of the data of our corpus using logistic regression is available in Otheguy, Zentella & Livert (2007).

1.5 The corpus and the envelope of variation

We limit the study to tokens of overt and null subject personal pronouns that occur with tokens of finite or tensed verbs, that is, verb forms inflected for person-number and tense-mood-aspect. The (many fewer) cases of subject personal pronouns occurring with infinitives and gerunds are not included in the study. Spanish speakers to be interviewed were selected according to stratification requirements based on several socio-demographic factors. The sample is thus well balanced with regard to (a) country of origin, (b) Latin American or NYC birth, (c) age of arrival in NYC, (d) years spent in NYC for those born in Latin America, (e) level of education (f) Spanish skills, (g) English skills, and (h) extent of use of Spanish with different types of interlocutors. Because of its large size and careful sampling, the corpus can claim to be representative of Spanish-speaking New York.

In the majority of environments where tensed verb tokens make personal references in Spanish, they appear in some cases with overt subject personal pronouns and in others with nulls. But some environments restrict verb tokens to appearing only or mostly with overts or, more frequently, only or mostly with nulls. In more technical terms, the majority of environments where tensed verbs are found making personal references are inside the envelope of pronominal variation and are eligible for a study of overt-null alternation such as the one we

present here, but other environments are outside the envelope of variation and are ineligible for such a study. We consider as being inside the envelope of pronominal variation and eligible for the study all tensed verb tokens whose subjects are animate beings that (a) occur in the corpus with an overt pronoun in an environment where a null could likely have been used, or (b) occur in the corpus with a null pronoun in an environment where an overt could likely have been used. All other verb tokens are excluded.

For example, we place inside the envelope of variation and include in our study the two verbs in *Quiero que él venga* 'I want him to come,' on the grounds that we could have found *Yo quiero que venga.* That is, the verb *quiero*, occurring with a null, could have been *yo quiero*, with an overt, while the verb *él venga*, with an overt, could have been *venga*, with a null. The 60,000-plus verb tokens of the corpus all occur in environments of these two types, where the subject personal pronoun is found in its overt form but could have been null, or is found in its null form but could have been overt.

We regard as outside the envelope of variation all verb tokens making references to inanimates, such as *Se rompió* 'It broke' and all tokens found with subject NPs that are not personal pronouns, such as *Carlos canta muy bien* 'Charles sings very well' or *Se trabaja mucho en NYC* 'Hard work is done in NYC.' These exclusions are on the grounds that the counterpart utterances with an overt pronoun occur very rarely in our corpus (e.g., *Ella se rompió*, referring to a wall or *Carlos él canta muy bien*); and on the grounds that, in more general terms, such utterances are exceedingly rare as well in natural speech. We also regard as ineligible for the study all meteorological environments like *No nieva mucho en Nueva York* 'It doesn't snow much in NYC,' on the grounds that they are not personal references, and that sentences like *El no nieva mucho en NYC* or *Ello no nieva mucho en NYC*, which can perhaps occasionally be found in at least some varieties of Spanish, are infrequent in general and very rare in our corpus. Similarly, we regard as outside the envelope of pronominal variation verbs found in subject-headed relative clauses such as *Él hombre <u>que le habló</u> de eso a tu padre se fue ayer* 'The man who spoke to your father about that left yesterday' on the grounds that neither in Spanish in general nor in the corpus in particular does one find in very high numbers sentences like *El hombre <u>que él le habló</u> de eso a tu padre se fue ayer.*[4]

A fuller analysis of the exact limits of the envelope of pronominal variation with tensed verbs in Spanish is beyond the scope of this paper. Further details on the criteria used for distinguishing between eligible verbs and ineligible verbs are available in Otheguy & Zentella (2007) and in Otheguy, Zentella & Livert (2007).

4. For an analysis of resumptive pronouns of this type, see Bentivoglio (2004).

1.6 The overt pronoun rate in the corpus as a whole

A brief explanation of the measure that will be used for the dependent variable is in order before proceeding. Results will be expressed in terms of the 'overt pronoun rate' or the 'pronoun rate.' The pronoun rate is the percentage of eligible verb tokens in the corpus (that is, the percentage of verb tokens inside the envelope of variation) that appear with an overt pronoun. Table 1 shows that the pronoun rate for the corpus as a whole is 33.5 percent, that is, 33.5 percent of verb tokens in the corpus appear with an overt pronoun whereas 66.5 percent of tokens appear with a null.

Table 1. Percent overt pronouns in the corpus

	N verbs	**Percent**
Verbs with Overt	21456	33.5
Verbs with Null	42505	66.5
Total	63961	100

We will see presently that the overall pronoun rate of 33.5 percent varies considerably between generational subsamples.

2. Generational differences

2.1 The pronoun rate in newcomers and in the New York raised (NYR)

We compare the group that is least exposed to the New York environment, newcomers, to the most exposed, the second-generation NYR, as these groups were defined above. We first raise the question whether the average pronoun rate is similar in the speakers from these two apparent-time generations.[5] The results are in Tables 2a, 2b and 2c; the first table includes all relevant consultants; the next two cover each of the regions separately.

5. We present here a cross-sectional, not a longitudinal study. The passage of one generation here is in apparent, not real, time. Our first-generation newcomers are not the actual parents of our second-generation NYR. But the Spanish usage of our newcomers is correctly assumed to be a good proxy for the Spanish that was brought to the City by the parents of our NYR when they were newcomers to the City. Even though our dependent variable is a percentage (pronoun rate), we employ ANOVAs to examine differences between groups as this rate is normally distributed in the Otheguy – Zentella corpus.

Table 2a. Average percent overt pronoun by generation (ANOVA)

	Avg % overt	N speakers	N verbs
Newcomers	30	39	17,130
NYR	38	28	12,101

F (1, 65) = 8.59, p < 0.005
Newcomers: Age of Arrival > 16 & Years NYC < 6
NYR: Age of arrival < 3.1

The table shows that, in the course of one apparent-time generation the pronoun rate has experienced a sharp, statistically significant increase; it is on average 30 percent among newcomers but 38 percent among the NYR.

The next two tables show that the pronoun rate increase takes place among the NYR with origins in both regional groups, but that the increase is much clearer among those with origin in the Mainland. Table 2b shows that the increase among Caribbean NYR is of six percentage points and of marginal statistical significance (p = 0.085). In contrast, Table 2c shows that the increase among Mainlander NYR is nine percentage points and clearly statistically significant (p < .005).

Table 2b. Average percent overt pronoun by generation, Caribbeans (ANOVA)

	Avg % overt	N speakers	N verbs
Caribbean Newcomers	36	19	8,844
Caribbean NYR	42	14	5,246

F (1, 31) = 3.17, p < 0.085
Newcomers: Age of Arrival > 16 & Years NYC < 6
NYR: Age of arrival < 3.1

Table 2c. Average percent overt pronoun by generation, Mainlanders (ANOVA)

	Avg % overt	N speakers	N verbs
Mainland Newcomers	24	20	8,286
Mainland NYR	33	14	6,855

F (1, 32) = 8.89, p < 0.005
Newcomers: Age of Arrival > 16 & Years NYC < 6
NYR: Age of arrival < 3.1

Tables 2b and 2c, displayed to show differential increases in the two regions, should also be read as documenting the general fact that Caribbeans, regardless of generation, show higher overt rates than Mainlanders. Among newcomers, whereas the rate for the Caribbean is 36 percent, it is only 24 percent for the Mainland, for a cross-regional spread of 12 percentage points. Among NYR, whereas the rate

for the Caribbean is 42 percent, it is only 33 percent for the Mainland, for a cross-regional spread of nine percentage points. (The cross-regional spread decreases because the cross-generational increase is higher in one regional group than in the other, as we have seen).

2.2 The overt pronoun rate and the variable Person by Region and Generation

Table 3 compares the pronoun rate for the variable Person for newcomers and NYR. Each panel in the table gives results for a cross-tabulation between Person and the pronoun rate. For each panel, we list the number of verbs on which the cross-tabulation was run (N verbs), the value of the chi-square (χ^2), and the probability value for establishing statistical significance (p). The table shows no value for second-person plural (2PL) because our Latin American consultants have no such verb endings (e.g. *cantáis*, 'you plural sing') in their verbal paradigm. The table marks off with solid lines those adjacent factors that are statistically significantly different from each other. Because we know that the regions differ with regard to this variable, the table shows the cross-generational comparison twice, once for Caribbeans and once for Mainlanders.

Table 3. Percent overt pronoun by Person & Generation, by Region (Cross-tabulations).

	Caribbean				**Mainland**			
	Newcomers		**NYR**		**Newcomers**		**NYR**	
	2SG	57	2SG	66	3SG	38	3SG	45
	1SG	38	1SG	51	1SG	28	1SG	36
	3SG	37	3SG	49	2SG	13	2SG	27
	1PL	18	1PL	19	3PL	12	3PL	16
	3PL	17	3PL	17	1PL	8	1PL	7
N verbs	8837		5236		8285		6853	
χ^2	484.63		480.96		445.67		427.43	
p	<0.001		<0.001		<0.001		<0.001	

The table shows, almost without exception, increases in the pronoun rate in all of the persons of the verb in both regions when we compare the newcomer to the NYR generation. And, with one important exception that we shall discuss presently, the table shows that, as the rates go up, the ranking of factors remains the same. The order [2SG – 1SG – 3SG] that is found in Caribbean newcomers is still found in Caribbean NYR. The order [3SG – 1SG – 2SG] that is found in Mainland newcomers is still found in Mainland NYR. And in both regions, the

plural forms remain the least likely to trigger overt use of the pronoun for both newcomers and NYR.

The table shows that an important qualification must be made to the generalization regarding cross-generational persistence in the order of the persons of the verb associated with overt pronouns. A change has taken place in the strength of 2SG among Mainland NYR. It no longer shares third place with the 3PL, from which it differed by only one percentage point (13 percent 2SG vs. 12 percent 3PL). Now 2SG among the NYR stands alone, having distinguished itself statistically from the 3PL by 11 percentage points (27 percent 2SG vs. 16 percent 3PL). Thus, 2SG among the Mainland NYR occurs at a rate similar to that of 1SG among Mainland newcomers, which occupied second place in that group. This change in the relative importance of the 2SG person is also reflected in the regional percentage-point spread. In the newcomer stage, the table shows a 44 percentage-point spread between the regions (Caribbeans 57 percent vs. Mainlanders 13 percent). But in the NYR stage, this regional difference has been reduced to 39 percentage-points (Caribbeans 66 percent vs. Mainlanders 27 percent).

One further consideration adds an important nuance to the general pattern of rates going up in the second generation with factor rankings remaining mostly the same. The percentage-point increases in the pronoun rate of the second generation are not uniform across the different persons of the verb, but appear to show a distinct pattern, as shown in Table 4.

Table 4. Pronoun rate percentage-point increases among NYR by Person and generation

	NYR increase	
	Caribbeans	**Mainland**
1SG	13	8
2SG	9	14
3SG	12	7

Table 4 shows that the percentage-point increases for the singular endings in the two regions are mirror images of each other. The forms that increase more among Caribbeans increase less among Mainlanders; the form that increases more among Mainlanders increases less among Caribbeans. The 1SG and 3SG person inflections show higher increases for Caribbeans (13 and 12 points respectively) and a lower increase for the Mainlanders (eight and seven points). The 2SG shows a smaller increase among Caribbeans (nine points) and a higher increase among Mainlanders (14 points).

2.3 The overt pronoun rate and the other independent variables

For the independent variables Continuity, Clause, and Lexical, which for reasons of space we summarize without displaying results in tables, the comparison between the generations yields a clear result. There are pronoun rate increases from newcomers to NYR in connection to all these variables, but no changes in factor orders, which remain the same among NYR as among newcomers. The NYR, like the newcomers, use more overts when the verb's subject is different from that of the previous verb than when it is the same; more overts in main and subordinate clauses than in coordinate ones; and more overts with verbs whose lexical meaning is mental-estimative than with verbs whose lexical meaning is action-external.

2.4 The overt pronoun rate in the different generations: Summary

The results of the bivariate comparisons between newcomers and NYR allow us to arrive at the following generalizations:

a. There is a sharp, statistically significant increase in the pronoun rate in the second generation. This cross-generational rate increase is noticeable among both NYR with origins in the Caribbean and those with origins in the Mainland, but is greater among Mainlanders.
b. For all four of the independent variables under consideration, there is an increase in pronoun rates in the second generation when compared to first-generation newcomers.
c. For three of the independent variables (Continuity, Clause, Lexical) the order of conditioning factors is the same in second generation speakers as in first-generation newcomers.
d. For one variable, Person, there is one cross-generational change in one region in the ranking of conditioning factors. Among Mainlanders, the 2SG Person inflection is, in a sense, higher in the scale among Mainland NYR than among Mainland newcomers; it is in third place by itself among NYR while it was tied for third among the newcomers.
e. The change registered among NYR with respect to the 2SG person relates not only to ranking but also to percentage points. Because NYR from the Mainland have increased their pronoun rate with 2SG persons more than have NYR from the Caribbean, the cross-regional spread is now less among NYR than it was among newcomers.
f. For the variable Person, the increases in overt pronoun use with singular verb forms among NYR from the two regions are mirror images of each other. The persons of the verb that show the largest cross-generational increases among

NYR from the Caribbean show the smallest increases among NYR from the Mainland. Conversely, the person of the verb that shows the largest cross-generational increase among Mainlanders shows the smallest increase among Caribbeans.

3. The overt pronoun rate in the different generations: An attempt at explanation

3.1 Language contact

The most reasonable interpretation for the increase in the occurrence rate of overt pronouns in the second generation is that Spanish usage in NYC, at least with respect to this feature, is undergoing a contact-induced change in the direction of English usage. Relevant considerations for this conclusion are that subject pronouns in English and Latin American Spanish are similar with regard to both their paradigms and their basic syntax. In both languages there are five distinct subject personal pronoun forms, and in both languages the overwhelming majority of subject pronouns are preposed to the verb. These similarities, which are enough to allow bilinguals to establish a parallel between the subject pronouns in the two languages, are accompanied by a well-known difference, namely that the languages differ with respect to the frequency and conditioning factors of the occurrences of these forms (for a study of nulls in English, see Cote 1996). Given these similarities and this large difference; given that, in general, languages tend to influence each other in bilingual settings (Weinreich 1953); that contact influences tend to flow from locally socially dominant to locally socially subordinate languages; that English and Spanish clearly occupy precisely these roles in NYC; and finally that the direction of lexical borrowing in NYC is primarily from English to Spanish while that of language attrition and shift is from Spanish to English, there is every reason to believe that the increase in subject pronoun rates among NYR is due to their adapting their usage of Spanish pronouns to that of the equivalent pronouns in English.

This conclusion is bolstered by the fact that, in our corpus, the pronoun rate correlates with a number of socio-demographic measures obtained from a questionnaire given to all our consultants, all of which point in the direction of the strong role of English in the Spanish pronoun rate increases, as shown in Table 5:

The table shows that, when the entire corpus is taken into account, the pronoun rate is positively correlated with years spent in the City (more years, more pronouns) and knowledge of English (better self-report of English, more pronouns), but inversely correlated with age of arrival (older arrival, fewer pronouns), knowledge of

Table 5. Pronoun rate and socio-demographic variables (Pearson correlations)

	N speakers	r
Years in New York	142	0.28**
English skills	141	0.24**
Age of arrival	142	−0.16*
Spanish skills	141	−0.19*
Spanish with siblings	125	−0.25**
Spanish with father	111	−0.20*
Spanish with workmates	126	0.20*

* = $p < 0.05$
** = $p < 0.01$

Spanish (better self-report of Spanish, fewer pronouns) and frequent use of Spanish with siblings and with father (more use of Spanish, fewer pronouns).

The correlations in the table point in the direction of English influence on Spanish usage, because they show that the individuals who have lived longer in a city where English is the dominant language, and the ones who feel most confident in their abilities in English tend to be the ones who use more overt pronouns in Spanish, whereas the ones who have been in the City less time and feel less confident in their English tend to use fewer pronouns in Spanish. In a set of secondary correlations that also associate the pronoun rate to English, we find that having arrived at an older age, knowing more Spanish, and using more Spanish with specific interlocutors, which suggests proportionately less use of English, leads to fewer Spanish overt pronouns. Only the last correlation in the table (more use of Spanish with workmates, more pronouns) lacks an unequivocal interpretation in support of the role of English in the increased use of overt pronouns in Spanish in New York.

3.2 Persistence in the second generation

The finding that the change in rates has taken place in the course of one generation without (with the exception of the variable Person) any accompanying change in the order of conditioning factors, tells us that second-generation speakers of Spanish continue to be guided in their use of subject pronouns, for the most part, by the same considerations as first-generation newcomers. Like the group of speakers that brought Spanish to NYC, New York bilinguals of both regions favor overts more in switch- than in same-reference contexts, more in clauses other than coordinates, and more in verbs that name mental-estimative events. Moreover, Caribbeans continue to favor overts especially for verbs with 2SG inflections, and Mainlanders for verbs with 3SG inflections.

3.3 Dialect contact

The larger generational increase in the pronoun rate among Mainlanders when compared to Caribbeans (cf. Tables 2b and 2c) is very likely due to the fact that Mainlanders are exposed in New York City to two high-pronoun forms of out-group speech. These two forms are English, where subject pronouns are usually overt, except in very specific environments, and Caribbean Spanish, where the rates are much higher than in the Mainland (Lipski 1994: 241, López-Morales 1992:137). In contrast, Caribbeans in NYC are exposed to only one high-pronoun out-group form of speech, namely English. A reasonable interpretation is that while the overall increase in rates is evidence of the fact that all Latinos in NYC are subject to language contact pressures, the greater increase among Mainlanders is evidence that the pressures on Mainlanders involve both language and dialect contact.[6]

Dialect contact is also the most reasonable interpretation of the cross-generational change in the ranking of conditioning factors for the variable Person in Mainlanders, among whom 2SG registered a movement upwards in the rank for the NYR. Given that among newcomers the 2SG inflection was a higher ranked factor among Caribbeans than among Mainlanders, this cross-generational change suggests movement by Mainlanders in the second generation toward a Caribbean usage pattern. In related points, the general reduction in the course of one generation of the cross-regional percentage-point spread, from 12 to 9 percentage points, which we discussed above with relation to Tables 2b and 2c, and the specific reduction of the cross-regional percentage-point spread in the 2SG, from 44 to 39 percentage points, which can be seen in Table 3, are also indicative of leveling in the second generation. (In Table 3, the 2SG cross-regional percentage-point difference among newcomers is 57 – 13 = 44; among the NYR it is 66 – 27 = 39).

Finally, dialect contact is also the most reasonable interpretation of the mirror-image pattern of rate increases in the variable Person, under which NYR from the Caribbean experience larger increases in pronoun rates precisely in the two persons of the verb that were higher-ranked among newcomers from the Mainland, while NYR from the Mainland experience a larger increase precisely in the person of the verb that was highest ranked among newcomers from the

6. Our reasoning here does not contemplate intra-Caribbean influences because even though some Caribbean groups do show higher rates than others (Dominicans have higher rates than Puerto Ricans and Cubans), these intra-Caribbean differences do not reach the clear levels of statistical significance that one finds when Caribbeans are compared to Mainlanders. It seems more reasonable, then, to think that the locus of influence is the regional group, and not the national units within it.

Caribbean. This pattern suggests that the NYR generation is leveling out its regional differences, changing its pronominal usage in a trend of mutual accommodation. A demonstration of this mutual leveling, and of the tilt towards Caribbean usage among Mainlanders with regard to the rank change of the 2SG, based on a multivariate analysis, is available in Otheguy, Zentella & Livert (2007).

4. Summary and conclusions

The comparison between the generations of Spanish-speakers in NYC shows four clear tendencies. First, the NYR use more overt pronouns than newcomers in all grammatical discourse environments studied in the present work, most likely due to the influence of the extremely high use of overt pronouns that is found in the other language of these bilinguals, namely English. Second, this increase in pronoun rates is more marked among NYR from the Mainland, most likely because they are adapting not only to English usage, but to Caribbean usage as well. Third, the ranking of factors in the Person variable shows the growing importance of 2SG inflections among Mainlanders, most likely as a manifestation of accommodation to Caribbean usage. Fourth, there is a mirror-image pattern to the cross-generational rate increase, indicating that the accommodating trend, while primarily in the direction of the Caribbean, is to some extent mutual between NYR of both regions.

The Spanish spoken by Latinos in New York City, then, at least with regard to the feature under study, is very similar that of the different regions of Latin America where they have come from, while showing, especially in the second generation, clear changes in usage due to language and dialect contact. The regions of origin resemble each other considerably on most fronts, but differ clearly in the motivation they assign to the different person-numbers of the verb for the choice between overt and null subject pronouns. This mostly similar but, in one respect, regionally differentiated Spanish, is marked by clear differences of usage in the second generation. (Differences in usage may also arise within members of the first generation who are not newcomers, that is, those who have lived in NYC for more than five years, but their speech has not been analyzed yet as of this writing.) The innovative usage is noticeable in all-around increases in occurrence rates, due to the influence of English, and in changes with regard to the variable Person. The Person changes represent, for the most part, adaptations by NYR from Colombia, Ecuador and Mexico to the usage of the Caribbean, but represent as well smaller adaptations by Cubans, Dominicans, and Puerto Ricans to the usage of the Latin American Mainland.

References

Bayley, R. & Pease-Alvarez, L. 1997. Null pronoun variation in Mexican-descent children's narrative discourse. *Language Variation and Change* 9: 349–371.

Bentivoglio, P. 1987. *Los sujetos pronominales de primera persona en el habla de Caracas*. Caracas: Universidad Central de Venezuela.

Bentivoglio, P. 2004. Las construcciones 'de retoma' en las cláusulas relativas: Un análisis variacionista. In *Lengua, variación y contexto: Estudios dedicados a Humberto López-Morales,* 507–520. Madrid: Arco Libros.

Cameron, R. 1993. Ambiguous agreement, functional compensation, and non-specific *tú* in the Spanish of San Juan, Puerto Rico and Madrid, Spain. *Language Variation and Change* 5: 305–334.

Cameron, R. 1995. The scope and limits of switch reference as a constraint on pronominal subject expression. *Hispanic Linguistics* 6/7: 1–27.

Cote, S. 1996. *Grammatical and Discourse Properties of Null Arguments in English.* PhD dissertation, University of Pennsylvania.

Enríquez, E. 1984. *El pronombre personal sujeto en la lengua española hablada en Madrid.* Madrid: Consejo Superior de Investigaciones Científicas.

Lapidus, N. & Otheguy, R. 2005a. Overt nonspecific *ellos* in Spanish in New York. *Spanish in Context* 2(2): 157–174.

Lapidus, N. & Otheguy, R. 2005b. Contact induced change? Overt nonspecific *Ellos* in Spanish in New York. In *Selected Proceedings of the Second Workshop on Spanish Sociolinguistics*, L. Sayahi & M. Westmoreland (eds), 67–75. Somerville MA: Cascadilla Proceedings Project.

Lipski, J. 1994. *Latin American Spanish.* London: Longman.

Lipski, J. 1996. Patterns of pronominal evolution in Cuban-American bilinguals. In *Spanish in contact,* A. Roca & J.B. Jensen (eds), 159–186. Somerville MA: Cascadilla Press.

López-Morales, H. 1992. *El español del Caribe.* Madrid: Editorial MAPFRE.

Montrul, S. 2004. Subject and object expression in Spanish heritage speakers: A case of morphosyntactic convergence. *Bilingualism: Language and Cognition* 7(2): 125–142.

Otheguy, R. & Zentella, A. C. 2007. Apuntes preliminares sobre el contacto lingüístico y dialectal en el uso pronominal del español en Nueva York. In *Spanish in Contact: Policy, Social and Linguistic Inquiries*, K. Potowski & R. Cameron (eds), 275–296. Amsterdam: John Benjamins.

Otheguy, R., Zentella, A. C. & Livert, D. 2007. Language and dialect contact in Spanish in New York: Toward the formation of a speech community. *Language* 83: 770–802.

Shin, N. L. & Otheguy, R. 2009. Shifting sensitivity to continuity of reference. *Español en Estados Unidos y otros contextos de contacto: Sociolingüística, ideología y pedagogía*, M. Lacorte & J. Leeman (eds), 111–136. Madrid & Frankfurt: Iberoamericana & Vervuert.

Silva-Corvalán, C. 1982. Subject expression and placement in Mexican American Spanish. *Spanish in the United States: Sociolinguistic Aspects,* J. Amastae & L. Elías-Olivares (eds.), 93–120. Cambridge: CUP.

Silva-Corvalán, C . 1994. *Language Contact and Change: Spanish in Los Angeles*. Oxford: OUP.

Silva-Corvalán, C. 1997a. Avances en el estudio de la variación sintáctica: La expresión del sujeto. *Cuadernos del Sur. Letras, Homenaje a Beatriz Fontanella de Weinberg* 27: 35–49.

Silva-Corvalán, C. 1997b. Variación Sintáctica en el discuros oral: Problemas metodologicos. In *Trabajos de Sociolingüística Hispánica*, F. Moreno-Fernández (ed.). Madrid: Universidad de Alcalá de Henares.

Silva-Corvalán, C. 1998. On borrowing as a mechanism of syntactic change. In *Romance Linguistics: Theoretical Perspectives,* A. Schwegler, B. Tranel & M. Uribe-Etxebarria (eds.), 225–246. Amsterdam: John Benjamins.

Toribio, A. J. 1994. Dialectal variation in the licensing of null referential and expletive subjects. In *Aspects of Romance Linguistics: Selected Papers from the Linguistic Symposium on Romance Languages XXIV,* C. Parodi, C. Quicoli, M. Saltarelli & M.L. Zubizarreta (eds.), 409–432. Washington DC: Georgetown University Press.

Toribio, A. J. 2000. Setting parametric limits on dialectal variation in Spanish. *Lingua* 10: 315–341.

Toribio, A. J. 2004. Convergence as an optimization strategy in bilingual speech: Evidence from code-switching. *Bilingualism: Language and Cognition* 7(2): 1–9.

Weinreich, U. 1953[1964]. *Languages in Contact. Findings and Problems* [Publications of the Linguistic Circle of New York 1]. The Hague: Mouton.

Zapata, G., Sánchez, L. & Toribio, A. J. 2005. Contact and contracting Spanish. *International Journal of Bilingualism* 9: 377–396.

Personal pronoun variation in language contact

Estonian in the United States

Piibi-Kai Kivik
Indiana University, Bloomington

The paper investigates variation in the form of personal pronouns in the informal speech of Estonians living in the United States (N = 23). VARBRUL analysis determined the factors influencing the variation of long and short form of personal pronoun and zero vs. pronominal subject. Three groups of speakers differed significantly: the late bilingual older WWII refugees, the early bilingual younger WWII refugees and the late bilingual recent immigrants. All speakers had maintained the functional long/short variation. The older refugees preferred long forms, possibly indicating a change in the monolingual community. The early bilingual speakers preferred overt pronouns, suggesting a language contact effect. The age of immigration, extent of education in L1 and L1/L2 use in networks appeared to correlate with patterns of pronoun use.

1. Introduction[1]

This paper presents results from a small-scale corpus study of variation in Estonian personal pronouns in a language-contact situation with English. Estonian, like many other languages, has two sets of personal pronoun forms, e.g., *mina* – long (L) and *ma* – short (S) denoting first person singular, 'I'. The full and reduced variants have been linked in traditional grammars to prosody, and there is functional variation with the long form used in an accented position. Studies of corpus data (e.g., Pajusalu 1996, 1997; Pool 1999; Kaiser 2003) have described variation

1. I would like to thank Dr. Julie Auger for her advice and help with this project, the anonymous reviewer for this volume and the LCTG audience for their comments. The remaining faults are my responsibility. My travel to the LCTG conference was partly funded by a grant from Indiana University Russian and East European Institute. I am grateful to everyone who helped with data collection, first of all to the participants in the study.

constrained by other factors or apparently unconstrained in some forms. Estonian 1st and 2nd person subjects can also be realized with zero pronouns.

The pronoun system with full and reduced forms (see e.g., Siewierska 2003 for a typological analysis of such systems) presents an interesting case for the study of language contact. Pronoun forms would not be expected to be easily affected, but a variable system itself could be subject to changes. Pronoun use is related to syntactic, pragmatic and discourse constraints, which are more resilient than lexicon, but long-term or intensive contact can have a restructuring effect. English, the contact language in this study, does not have comparable morphological variants for personal pronouns, and does not allow zero subjects. Aspects of the Estonian system have similarities with languages with strong-weak (clitic) pronoun series and null-subject languages, which have both been shown to be affected in various ways by contact with English (e.g., Bavin 1989; King 1989; Satterfield 2003, Otheguy et al. this volume).

The corpus in this study consisted of sociolinguistic interviews with first-generation Estonian immigrants or long-term sojourners in the United States. Attempts at a systematic study of Estonian in immigrant communities have only recently begun (e.g., the papers in Lindström 1998; Klaas 2002) with the exception of some earlier work (e.g., Oksaar 1972; Roos 1980; Raag 1982, 1983), and have focused mainly on lexical, morphological (inflectional) and phonological, to lesser extent syntactic aspects (Raag 1985; Lehiste & Kitching 1998; Nemvalts 1998; Maandi 1989), see also the overview of Swedish Estonian by Raag (1991). Although small-scale, this is the first corpus study to look at pronoun use by American Estonian speakers.

2. Background

2.1 Overview of the Estonian personal pronouns

Estonian personal pronouns, like all nominals in the language, have inflections for case and number. There are three abstract cases (nominative, genitive and partitive) and 11 adverbial cases (Viitso 2003: 32). Personal pronouns have parallel short and long forms in most, but not all case-forms of the inflectional system (see Appendix 1).

In Standard Estonian, the variation of long and short forms is similar to the one observed, for example, with Dutch feminine full and reduced forms, where full forms occur in case of emphasis, but unstressed use is possible as well, except for contrast and coordination (e.g., Kaiser & Trueswell 2004). *Short* and *long pronoun* are the terms used in this paper following the tradition in Estonian

grammars. While the nominative (subject) and genitive short forms in Estonian, which have a CV structure, exhibit clitic-like behavior, some Estonian short forms are reduced, but not clitic-like, e.g., the allative *minule* (L) – *mulle* (S) 'to me' (L refers to *long*, S to *short* form in this paper).

For the clear-cut cases where the long forms are specified for stressed pronouns, the variation in personal pronouns can be described as *referential choice* (e.g., Grüning & Kibrik 2005). It has been noted about languages with reduced and non-reduced pronouns that the reduced ones are for anaphoricity, and the non-reduced ones "have focus functions" (de Hoop 2003: 159). Normally, the deictic pronoun forms are *long* in Standard Estonian, but the long form can also be used anaphorically, and it does not always have to bear stress. The accented long form sometimes corresponds to left-dislocation in English. De Hoop (2003) claims that the stressed (accented) pronouns in English are indicative of contrast between two situations in the discourse. This certainly is the case with Estonian long forms, as opposition requires the use of *long* in Estonian even without the expected accentuation (Pool 1999). The contrast signalled can be implicit, as in switch reference. Pajusalu (1996, 1997, 2005), Kaiser and Hiietam (2003), Kaiser (2004) have pointed to the contrast-indicating function in Estonian 3rd person long pronoun. In sum, there is no one-to-one correspondence between English stressed pronouns and Estonian long forms. When stressed pronouns in English usually correspond to *long* in Estonian, not all *long* Estonian pronouns correspond to the stressed ones in English.

The first corpus-based investigation of the Estonian short and long forms was by Pool (1999). Her study analyzed the use of the first and second person singular pronouns in oral conversations, written drama texts, and texts from the Corpus of Estonian Literary Language. The selection of the pronoun form was found to depend on information-structure of the sentence (prominence) and case form of the word (Pool 1999: 179). The *long* form was used for important (new) information and opposition. As to case, the short form was preferred in the nominative (subject case) as well as allative and adessive (*minule (L)/mulle(S)* 'to me', *minul(L) /mul(S) on* 'I have'). Allative and adessive cases are used to refer to the experiencer and possessor, i.e., the semantic subject in the corresponding constructions. Pool's findings are in line with the relationship between argument prominence and distribution of reduced pronominals, suggested by Bresnan (1998, cited in Siewierska 2003) and typologically investigated by Siewierska (2003): reduced pronominals are most common with subjects and decrease with the obliqueness of arguments. However, in Pool's study, some cases of the use of unstressed long forms remained unaccounted for. Similarly, the choice of *long* vs. *short* for non-prominent genitives (*minu(L)/ mu(S) isa* 'my father') was assumed to be in free variation.

The Estonian subjects can also be realized with the zero pronominal. According to the traditional grammar, this is only the option in 1st and 2nd person. Estonian appears to be a *mixed null-subject* language, a term used by Vainikka and Levy (1999) in their account of a similar system in Standard Finnish and Hebrew. While it is true that main clauses in isolated sentences are not grammatical with zero subject, as in (1), third person subjects can be dropped in certain narrative registers when the referent has been sufficiently identified (Lindström 2001; Keevallik 2003), possibly as a topic-drop or discourse ellipsis (cf., Lindseth 1995 for Russian and Slavic) or a register-specific feature akin to the English diary-style drop (Haegeman 2000).

(1) a. **Astus* *rongile*
step-PST.3SG train-onto
'(He/she) boarded the train'
b. *Astusin* *rongile*
step-PST.1SG train-onto
'I boarded the train'

Unlike in true pro-drop languages, 1st and 2nd person subject zeros are not the unmarked form in Estonian. 1st person drop is a stylistic feature in colloquial spoken narratives (Lindström 2001) on the one hand, and in semi-formal writing (narrative CV-s, reports, etc.) on the other, but not a consistent marker of register.

Discourse contexts for subject drop in colloquial spoken Estonian were specified by Duvallon and Chalvin (2004). 49% of the 2nd person singular verb forms and 18% of the first-person verbs were found to be realized with zero subjects. The syntactic context favoring zero subjects was found to be the presence of a preverbal direct object or another verbal complement. The XVS order, the inverted word order, is the order for about 50% of Estonian declarative clauses, SVO with a V2 preference is considered the basic word order (Ehala 2001; Erelt 2003). However, studies of spoken Estonian and conversation data have shown greater preference for SO and SV orders, as Lindström (2005) found that the occurrence of XVS in spoken non-narrative speech is only about 5%. Corpus studies have also documented 3rd person subject drop and verb-initial clauses in narratives (Lindström 2001; Keevallik 2003; Võik 1990, cited in Keevallik 2003).

2.2 Language contact research

In the extensive literature on language contact, the following factors have been found to influence L1 maintenance or attrition in expatriate speakers: time since immigration, age at onset of bilingualism, education, language contact and social

networks, L2 proficiency, and individual differences. Below, some findings involving the factors relevant for this study will be reviewed.

Length of immigration may not always be a factor. However, there is interaction between contact and time since immigration (de Bot et al. 1991; Köpke & Schmid 2002; Hutz 2002; Jarvis 2003).There are important distinctions between attrition in children and adults. Köpke and Schmid (2002: 10) observed that among speakers who were above the age of 12 when L1 input was reduced, "the amount of attrition was surprisingly low, even after many decades spent in an L2 environment." Similarly, education in the L1 can protect the L1 from attrition in an L2 environment, e.g., Pelc 2001 (cited in Köpke & Schmid 2002) found that for Greek-English bilinguals, years of general education in Greece (prior to their immigration and education in the United States) correlated with their linguistic performance in L1 Greek. Lehiste and Kitching (1998), who studied the use of object cases by Estonians in North America, found that the group of informants who had not attended school in Estonia differed from those who had, despite minor age differences. The former were unsure about the cases required for total object, whereas the latter behaved according to the monolingual norm.

Social networks, both primary and secondary, can have an effect on maintenance and attrition, although de Bot and Stoessel (2002) noted the lack of quantitative support for a direct relationship between social networks and language use. They also pointed to the recent change in the nature of social networks thanks to the emergence of the Internet. There can also be considerable individual differences in attrition for speakers of similar social characteristics (e.g., Altenberg 1991).

With regard to change in the L1 of expatriate bilinguals compared to the monolingual home-country community, the cohort effect of preserving an earlier state of the language has been described by e.g., de Bot, Gommand and Rossing (1991) for Dutch in France, Yağmur (2002) for Turkish in Australia. Preservation of dialect features has also been noted as a characteristic of exile Estonian communities. In addition to the physical isolation from the language community in Estonia, these speakers have not been subject to the leveling and standardization that took place in Estonia (papers in Lindström 1998; Keevallik 2003).

Internal processes of change can be accelerated in contact situations (e.g., Boyd and Andersson 1991; Ben-Rafael 2002; and Maandi 1989 for Estonian object case marking in Swedish Estonian). Lainio and Wande's (1994) variation study of the Finnish subject pronoun in the speech of Finns in Sweden found increased use of an overt pronoun where Standard Finnish requires zero, while overt subject pronoun has also become the norm in Colloquial Finland Finnish.

Changes in pronoun use may reflect a change in syntactic patterns. Syntax is known to be much more resistant to contact-induced changes than the lexicon

or the phonetic system (Hutz 2002; Sankoff 2002; Toribio 2001). The main effects described have been the simplification of syntactic patterns, e.g., Maher (1991) in her study of an enclave speech community Finnish in Minnesota, and restriction of word-order choices, over-generalizing the SVO order in contact with English. A longitudinal study of speakers of L1 German in the United States noted "higher frequency of SVO order in main clauses," despite very little syntactic change (Hutz 2002: 202). In Jarvis' (2003) case study, Aino, a Finnish-English bilingual in the United States, changed the word orders produced by native speakers in Finland to canonical SVO orders in an error-correction task. Satterfield (2003) showed that the bilingual Spanish-English speakers (acquired Spanish natively) did not maintain the monolinguals' distribution of non-overt and lexical subjects in their speech. Neither did they exhibit the monolinguals' rigid focus/contrastive distinctions with regard to pronominal interpretation. Heine and Kuteva (2005) provided a synopsis of research testifying to similar effects.

There is substantial evidence to show that immigrants in the USA are likely to activate use patterns that are marginal in their respective L1 by using them more frequently and extending them to novel contexts in cases where their L2, English, provides a convenient model (Heine & Kuteva 2005: 68). The criteria proposed by Heine and Kuteva (2005) for a this change from minor to major use pattern, from optional to obligatorily marked category were as follows:

1. Frequency (personal pronouns used at higher frequency).
2. Context extension (personal pronouns used in contexts not found in L1 but present in L2).
3. Change in meaning. "Originally serving pragmatically defined functions (presenting new or topical participants), the pronouns increasingly lose these functions and assume the syntactic functions of presenting pronominal subjects (or objects)." (p. 70)

3. Estonian pronoun variation and contact: Hypotheses

In a language contact situation where an L1 is spoken in an L2 environment, among the "sites of high attrition likelihood" (Preston, cited in Sharwood Smith 1989: 191) there are "unique items, distinctions in L1 that do not exist in L2" and "synonymous items." Seliger and Vago (1991) noted that the relationship where L2 unmarked forms replace a L1 marked form is the most likely to produce language attrition.

Estonian variable personal pronouns can be assumed to be prone to attrition in contact with English, especially as there are cases with no clear prosody-form

correspondence. The forms are synonymous, and in comparison with English, the long form appears morphologically marked. English, at the same time, has stressed and unstressed pronouns, which in many cases pattern similarly to the Estonian long and short form. The mixed null-subject Estonian can be assumed to undergo changes from minor to major use pattern as has been observed with the null subject languages, i.e., the SVO order with overt subjects will dominate.

The following hypotheses were put forward:

1. Variation of personal pronoun forms in the expatriate Estonian community differs from the patterns observed with monolingual Estonian speakers.
2. The more extensive the language contact, the less variation there is, with the short forms preferred.
3. Zero subjects are dispreferred by speakers with extensive English contact.

Additionally, the following questions were asked: are there differences between speakers in the language contact situation related to their background such as age of immigration and network characteristics? Are there differences between the recent immigrants and the WW II refugees?

In this paper, only the data from Estonian American speakers will be analyzed. Comparisons with the Estonian speakers in Estonia will be based on published studies. For space considerations, these data will not be reproduced here and the readers are referred to the sources.

4. Method

Data in the form of sociolinguistic interviews were collected from first-generation Estonian-speakers at three urban locations in the United States. The total sample from the three locations consisted of 23 speakers. Two had to be excluded from quantitative analysis due to the small amount of tokens.

4.1 Background of the sample

The majority of Estonians in North America today are WW II refugees who fled Soviet occupation, and their descendants. The escape was often very sudden and unplanned. Most perceived it as temporary in the beginning and hoped to return soon. Estonian exiles approached life in America with determination to resist Americanization (Raun 2004). One of the main goals of the exile community was the preservation of Estonian culture, identity and language.

Since Estonia re-gained independence in 1991 and people were once again free to travel, a number of Estonians have moved to the U.S., temporarily or permanently, for work, study or family reasons. This new group of immigrants tends to be more dispersed and not necessarily integrated into the existing refugee communities.

4.2 Participants

The 23 participants (21 included in the quantitative analysis) fall into three groups, which were constructed based on the type of immigration (WWII refugee or recent immigrant), age at immigration and whether or not secondary education was in Estonian, i.e. late and early bilinguals: Group 1, the *older* refugees: N = 10 (11); M = 4, F = 6 (7); Group 2, the *younger* refugees: N = 6; M = 3, F = 3; and Group 3, the *new* immigrants: N = 5 (6); M = 2 (3), F = 3. Tables 1 and 2 present the distribution of participants by groups, age, immigration data and level of education, more information is in Appendix 2. The age of the majority of the participants was above fifty, therefore age-grading should be considered when interpreting the results. For a few participants, their status as a member of *older* or *younger* group was not unambiguous. The decision was made based on whether the speaker had obtained secondary education (high-school) in Estonian or in English, and some members of the first and second group are quite close as to their age. Several *older* group members had graduated from the Estonian *Gymnasium* in Germany.

Table 1. Age and immigration data of the participants

	Group	N	Age	Age of arrival	Years in the U.S.
1	Older	10	68–75	16–31	44–55
2	Younger	6	55–67	3–15	50–53
3	New	7	(18)24–37	(14)20–29	3–8

Table 2. Educational background of the participants

	Group	Did not complete high school	High school/ Vocational school	Undergraduate degree	Graduate degree
1	Older	2	(1)	4	3
2	Younger	–	–	1	5
3	New	(1)	1	–	5

The participants were located by first approaching the central figures in the Estonian communities,[2] who facilitated recruiting at Estonian community meetings and suggested potential participants. Also, snowball sampling was used, as participants suggested other speakers to contact. Therefore, the sample is somewhat biased towards people who, first of all, had contacts with other Estonian speakers or were active in their communities, and, secondly, were interested in talking to the researcher. In the refugee population, these speakers were often well-educated, socially well-established and linguistically and culturally strongly maintenance-oriented. The *new* immigrants who agreed to participate were professionals, advanced graduate students or permanent U.S. residents, thus the sample is not representative of all the Estonians who have recently settled in the U.S.

4.3 Procedure

The interviews lasted from 45 minutes to over an hour and were conducted as an informal conversation with the researcher. They were designed to elicit life-stories about coming to America, about the activities in the Estonian community, memories from Estonia or impressions from recent trips. The recent immigrants were also asked about differences they noticed compared to life in Estonia. The speakers were not aware of the specific linguistic focus of the study. They were interviewed in their homes, in the community centers or, in case of two Group 3 speakers, in a university office.

5. Analysis

The tape-recorded interviews were transcribed and personal pronoun forms were coded for statistical analysis. The coding system included the independent variables of *long* and *short* pronouns and the dependent variables in the five factor groups as shown in Table 3.

For an additional analysis, all subject pronouns were coded as overt or zero. Differently from Pool's (1999) study, the Focus/Prominent factor group distinguished between pronouns that were either made to stand out (accent, focus position, particle) in the utterance (Focus factor), were explicitly contrasted with another NP (Contrast factor) or did not have any of these characteristics (Non-prominent).

---

2. The selection of the communities was opportunistic. This is a small-scale unfunded study and the sample size and sampling methods are the main limitations to generalizing beyond the corpus analyzed.

Table 3. Coding system (Examples are given in nominative or 1st person for brevity. CAPS denote accentuation.)

Factor	Example, notes of use
I Person	
1. First person singular	*ma / mina*
2. Second person singular	*sa / sina*
3. Third person singular	*ta / tema*
4. First person plural	*me / meie*
5. Second person plural	*te / teie*
6. Third person plural	*nad / nemad*
II Case	
1. Nominative	*ma / mina olen* 'I am' Subject case
2. Genitive	*mu / minu vend* 'my brother' , *mu / minu arvamus* 'my opinion', *minu / mu juures* 'at, near me' possessive pronoun, head noun of postposition, premodifier, direct object with perfective verb
3. Locative-inner	*mus /minus* 'in me', *must /minust* 'from (about) me', *musse / minusse* 'into me'
4. Allative	*mulle / minule* 'to me', *mulle / minule meeldib* 'I like' direction to, indirect object (some Oblique or Dative functions), notional subject with the verb *meeldima* 'like'
5. Adessive	*mul / minul on* 'I have' location at, notional subject with some experiencer phrases, e.g., possession
6. Other cases	*minult / mult* 'from me', *minuks / muks(?)* 'as/for me'
7. Comitative	*minuga / muga* 'with me'
III Focus/prominence	
1. Focus/prominence	*Mina küll tahan kohvi.* 'I for sure want coffee'. *Siis tuli TEMA.* 'And then came HE'.
2. No prominence	*Ma elasin Tallinnas.* 'I lived in Tallinn.'
3. Contrast	*Mina olen Tallinnast ja tema on Tartust.* 'I am from Tallinn and she is from Tartu.'
IV Gender	
V Speaker	

Care was taken to code for Contrast and Focus independently of the long form itself, which for a native speaker immediately signals prominence. Contrast and Focus as factor groups should be understood as non-technical uses of the terms and not equated with the respective syntactic positions.

Statistical analysis was performed with the Varbrul (GoldVarb 2) software, (Paolillo 2002; Young and Bayley 1996; Labov, to appear). Two separate analyses were performed. The independent variables in the first Varbrul analysis were

pronoun forms: *long (L)* or *short (S)*, 4472 tokens were included. In the second, subject pronoun analysis, the independent variables were overt pronoun or zero, 4184 tokens were included. In the initial L-S Varbrul runs, all pronouns were included. Later, to strengthen the model, the prominent pronouns (focus, contrast) were excluded so that variation due to other factors would emerge more clearly.

6. Results

6.1 Total sample

For the total sample, the Varbrul analysis confirmed the factors suggested by Pool (1999) as significant for the choice of *long/ short* forms: prominence and focus and contrast were the strongest predictors of *long* forms. Case form was also significant: the strongest preference for *long* was with inner locatives (.99), comitative (.96), and genitive (.97, p = .001), the weakest with nominative case (Varbrul values closer to 1 indicate preference for long forms). In addition, person and number also proved significant. The form most likely to be realized with *long* was third person plural (.74) and the one most likely to be realized with *short* was second person singular (.36, p = .001). Gender was not a significant factor group.

Zero subject personal pronouns. In the total corpus, the factors influencing the choice of overt pronoun over zero were that of person and individual speaker. Least preference for overt subject pronouns (Varbrul values closer to 1 indicate preference for overt) was found with 2nd person (.38), both overt and zero subjects were preferred with the 1st person (.50) and a very slight preference for *overt* was observed in case of 3rd person (.54). This result concurs with Duvallon and Chalvin's (2004), who found more zeros for second person than first person in monolingual corpus. The 1st and 2nd as opposed to 3rd person preferences for overt over zero appear to be the same as for long over short, suggesting salience/accessibility plays a role in both cases. Speaker differences will be discussed below.

6.2 Speaker groups

The individual speaker category was significant (Appendix 3 lists the Varbrul weights). Analysis after recoding the data for three speaker groups of older refugees (Group 1, late bilinguals), younger refugees (Group 2, early bilinguals) and new immigrants (Group 3) yielded the group factor as significant. The three groups proved to be statistically different in their preferences for long forms and overt subject pronouns.

Long vs. short. The three groups differed significantly in their preference for *long*. The distribution and Varbrul weights for the preference of *long* are shown in Table 4 (tokens of nonprominent pronouns). The Varbrul probabilities show that the younger refugees and the new immigrants had a similar pattern, whereas the *older* group differed. The *new* group had only a slight preference for short forms (.42), the *younger* group showed a preference for short forms (.35) and the *older* speakers preferred long forms (.73).

Table 4. Distribution of Long and Short in all nonprominent pronouns, preference for Long, speaker groups

Group	Long		Short		Total		Weight
	%	(N)	%	(N)	%	(N)	
Group 1: Older refugees	18	(218)	81	(955)	30	(1173)	.73
Group 2: Younger refugees	8	(165)	91	(1790)	50	(1955)	.35
Group 3: New immigrants	6	(45)	93	(691)	19	(736)	.42
Total sample	11	(428)	89	(3436)	100	(3864)	

Short forms would be expected to prevail in conversation. Both the new immigrants and the younger refugees showed this result, and their use was not suggestive of excessive or categorical use of *short*. The actual preference for *long* over *short* cannot be concluded from this analysis, as the speakers may also have differed in their use of zeros: the group with a preference for *long* might also have more zeros and thus just show more variation. Therefore, these results should be combined with those of the overt vs. zero analysis.

Overt vs. zero. With regard to the zero subjects, the group patterns were different from those observed for *long vs. short*. Here, the *older* refugees (Group 1) and *new* immigrants (Group 3) patterned together and the *younger* refugees (Group 2) differed (Table 5).

Table 5. Distribution of overt vs. zero subject pronouns, preference for overt

Group	Overt		Zero		Total		Weight
	%	(N)	%	(N)	%	(N)	
Group 1: Older refugees	80	(1171)	19	(284)	34	(1455)	.43
Group 2: Younger refugees	90	(1649)	9	(177)	43	(1826)	.62
Group 3: New immigrants	76	(694)	23	(209)	21	(903)	.38
Total sample	84	(3514)	16	(670)	100	(4184)	

Older refugees (.43) and recent immigrants (.38) indicated a slight preference for zeros whereas the younger refugee group's preference for overt subjects was significantly higher (.62).

Combining the results of the long-short and overt-zero analyses, the following can be observed. As the *older* and *younger* refugee groups differed in their use of zeros, the present *long-short* analysis only indicated overall reduction of variation and increase of *short* for the younger group, not necessarily *short* overtaking the functions of *long*.

At the same time, as the older refugee and new immigrant groups did not differ significantly in their preference for zeros, the *older* group's preference for *long* indicates that the *older* speakers had a different *long/short* usage pattern.

6.3 Comparison of results with Estonian speakers in Estonia

The distribution of 1st and 2nd person subject forms for the *older* refugees (Group 1) was most similar to the Estonian Conversation results (Pool, 1999), with slightly higher occurrence of long forms (15% compared to 11% in the Estonian Conversation). *Younger* refugees (Group 2) had a percentage of long forms closer to the Estonian Conversation, but they looked very different in their use of zeros, only 7% compared to 25% in Estonian Conversation. The *new* immigrants (Group 3), in turn, were very similar to the Estonian Conversation speakers with respect to the use of zeros. Surprisingly, they showed a lower frequency of long pronouns (3% vs. 11% in Estonian Conversation).

The *older* and the *younger* refugee group were similar with respect to their distribution of *long* vs. *short*, and they were also like the monolingual speakers regarding these variants. The *new* immigrants were different from the other American groups and also from the Estonian data, as their long forms were less frequent.[3]

With regard to zeros, the *older* and the *new* group were similar and they were also close to the monolinguals in that respect. The outliers here were the *younger* refugees, who differed from other American groups and the Estonian samples in their higher frequency of overt forms.

3. It is possible that Pool's sample included more older or conservative speakers compared to Group 3 in this study.

6.4 Genitive

Of the total genitives, 74% were long in conversations reported in Pool (1999). In my comparable data, 61% were long, indicating a higher preference for short genitives. In the same, 1st and 2nd person singular subsample, for non-prominent genitives, 39% were long for *younger* refugees, 62% were long for *older* refugees and 63% were long for *new* immigrants. The *older* refugees and *new* immigrants (late bilinguals) were very similar, with a preference for long genitives even if the pronoun was not prominent. The *younger* refugee group (early bilinguals) still preferred short forms, as with all other case-forms.

Example (2) illustrates the use of long genitive for focused as well as non-focused pronouns by a Group 3 speaker.

(2) Toomas' story of a woman with hand-grenade
1 *ja miks tal see granaat seal oli sis sellepärast*
'and why she-(S) had this grenade there, this is why'
2 *see ei ole üldse TEMA granaat*
'it is not her-(L) grenade at all'
3 *vaid see oli hoopis tema BOIfrendi granaat*
'but it was her-(L) boyfriend's grenade instead'
4 *ja tema boifrend töötas mingisuguses sõjaväeosas*
'and her-(L) boyfriend worked in some army unit'

In line 2, *tema* 'her' is contrasted with *tema boifrendi* 'her boyfriend's' in line 3, it is accentuated and thus prominent. In lines 3 and 4, however, there is no prominence or focus of any sort on *tema* 'her', but the long form is used anyway.

For the total sample, preference for genitive long forms was found to be in inverse relationship to preference for overt subjects.

7. Discussion

The main finding with regard to the whole corpus was that functional variation in the American Estonian sample in this study was well maintained, with (the universal) factors determining pronoun choice similar to those identified earlier for the monolinguals: prominence (defined as focusing or contrast) and case inflection. Thus the variable system was not significantly affected by contact with English in this first-generation immigrant corpus.

The two groups of first-generation refugees and the third group of new immigrants were all different in their variation patterns. The major findings were the similarities of the older refugee group and the recent immigrant group, i.e. the

two groups of late bilinguals, in their use of zero subjects and long non-prominent genitives. On both measures, the younger refugee group, the early bilinguals most likely to display language contact effects, differed from the other groups in the predictable direction of preferring short pronoun forms over long and zero. Overall, the use of long vs. short pronouns was not found to be clearly indicative of language contact. The older refugee group was found to show stronger preference for *long* forms compared to other groups, possibly as an indication of a maintained older pattern or style feature. While both the younger refugees and new immigrants preferred short forms, their reasons for doing so cannot be the same. The new immigrant speakers left their home-country only recently and were likely to represent newer monolingual trends, which may have been accelerated in the contact situation, thus their dispreference for long forms. The younger refugees, at the same time, were early bilinguals growing up in expatriate communities, whose linguistic choices were shaped by entirely different factors.

Any differences between the older refugees (strong maintainers) and new immigrants on the one hand, and the younger refugee group, on the other hand, would suggest a language contact and/or L1 attrition effect. Such a difference observed in this study was the greater preference for overt subject pronouns in the younger immigrant group. A comparison with Duvallon and Chalvin's (2004) study suggested that in the 1st and 2nd person the younger refugee group also used more overt pronouns than speakers in Estonia. The hypothesis of the restriction of the use of zero subjects with language contact was confirmed.

Incomplete acquisition, L1 education, age. The younger refugee group was initially constructed on the basis of the age of arrival in the U.S. (in mid-teens or younger), which typically also involved at least secondary school in English, i.e., lack of exposure to Estonian-language secondary or post-secondary education. The results of this study appear to support other studies (e.g., as cited by Köpke & Schmid 2002) that have pointed to the importance of L1 education for L1 maintenance.

The younger refugee group exhibited social characteristics more common to the second-generation immigrants and incomplete acquisition cannot be completely ruled out. The speech of the parents of this group and thus their own initial L1 acquisition was not likely affected by language contact. However, the fact that most, if not all, of the education and professional communication of this group has been in English suggests that their L1 never fully developed (cf. the discussion of late acquisition of discourse features in Extra & Verhoeven 1999: 36–38). The younger refugee group could be displaying a very colloquial style, the home use of their parents that had never been supplemented by other registers

Several factors need to be considered here. The preference for overt subjects could be related to increased processing load for bilinguals, as has been suggested

by Satterfield (2003) for Spanish immigrants in the United States. Indirectly, overt subjects can be related to the SVO word order, as Duvallon and Chalvin (2004) identified XVS word order as one of the contexts favoring subject drop in Estonian. The large number of overt subjects could thus suggest syntactic change in progress, which again could point to either processing difficulties or changes in the language system. The speakers rely heavily on SVO order, which in turn calls for the use of overt subjects in V-2 Estonian.

As SVO is one of the prevalent orders in Estonian, Heine and Kuteva's (2005) spread from a minor to major use pattern could be relevant here, perhaps more in terms of one of the major patterns spreading. In mixed-null subject Estonian, essentially a similar process to the null-subject languages can be observed in contact with English.

However, in the light of the Finnish data from Lainio and Wande (1994) and Estonian conversation data from Lindström (2005), increase in the use of short pronouns and SVO, respectively, could be a trend in Estonian. The internal change can be coinciding with the external influence and processing factors in contact situation. Further studies need to tease these issues apart.

Networks and contacts. The three groups all had central patterns and speakers who diverged from these. The divergent speakers were usually found to have patterns that placed them closer to another group that they also had close contacts with: Timo, the Group 3 outlier, had close contacts with older refugee in-laws, and his isolation from L1 peers made his speech resemble Group 1 patterns. Another "new immigrant" who had close contacts with the other groups, was Diana, who still had her Group 3 pattern, but her preference for overt subjects was much higher than for the other three "newer" immigrants. Diana's background resembled that of the Finnish woman Aino reported in Jarvis (2003). Diana's increased preference for overt subjects compared to the other new immigrants could also reflect a SVO preference that Jarvis observed about Aino.

Attitudes. The groups in the study differed in their attitudes towards maintenance. The new immigrants did not express worries about maintaining their language. They were confident that modern communication facilities kept them in close contact with Estonia and the language. Those who had firmly settled in the U.S., actively sought out opportunities to maintain their language and culture, such as sending their children to Estonian camps or participating in cultural events. However, they saw linguistic and cultural maintenance as an individual's free choice, differently from the refugees' strong orientation towards maintenance as a group. Clearly, it is the changed political reality behind these differences. Estonia is a free independent state with Estonian as its official language. For the

WWII refugees, their home-land was occupied and Estonian threatened by Russian, the *lingua franca* of the occupying state. Due to historical reasons, Estonian language has been an extremely important part if not the core of Estonian identity. These differences in attitudes could make an important difference in actual maintenance/ attrition patterns that should be considered for future studies of current immigrants.

For speakers in this study, attitudes towards assimilation and the new homeland appeared to correlate with the linguistic patterns observed, although there were too many interacting factors for a conclusive relationship. Ella in Group 1, who had the most zeros and most long forms, thus showing variation patterns most divergent from English, reported coming to the U.S. very unwillingly from her already established home in post-war Europe. She was also the oldest in the sample when immigrating, and her social networks consisted mainly of Estonian-speaking family and friends. The Group 2 speakers with patterns most suggestive of contact effects, such as Juhan and Ulla, had been mostly assimilated, as they were married to Americans and had well-established professional lives. At the same time, Arno, who shared the same family and professional characteristics, was much more Estonia-oriented and put a lot of effort into cultural maintenance. Individual differences between the patterns observed in speakers of similar social characteristics, such as Arno and Juhan, point to the importance of attitudes and active contacts with L1, especially in the context of L2 primary networks, for maintaining the pronoun variation patterns.

8. Conclusions

In this study of Estonian American speech, the functional variation in the personal pronoun system was observed to follow patterns very similar to those described in the monolingual corpus. The overall functional variation appeared to have been well maintained and was not considered a measure of language contact effects, at least not across all personal pronoun forms. Instead, the markedly different pattern of the older refugee group in their preference for long forms suggested a possible language change in the monolingual community. As a possible language contact effect, those speakers for whom contact with English was considered to be the most intensive, showed less distinction between genitives and other forms, preferring short forms in both, whereas the monolingual norm prefers long genitives in conversation.

The use of overt subjects in place of zeros was found to be a better indicator of language contact, as the speakers considered most likely to experience language

contact effects (education and networks in L2 English, limited L1 contacts) demonstrated higher preference for overt pronouns. The age of immigration, extent of education in L1 as well as L1/L2 use in networks appeared to correlate with changes in pronoun use, whereas length of stay did not.

The study observed stability of the functional variation in long-term immigrants with strong maintenance orientation, relative stability in the overall corpus. Discourse features of pronoun usage had been preserved, especially by maintenance-oriented educated speakers despite being cut off from home country for 50 years. Increased preference for overt subjects for speakers who were considered to show language contact effects was suggested to result form a syntactic shift to stronger SVO preference, or from the reduction of stylistic options.

The study hoped to contribute both to the study of variation and pronominal reference in Estonian and the study of Estonian in expatriate communities. Further studies both in the expatriate and home-country communities are needed to determine the details of language contact effects in the use of the variable pronouns and to specify the extra-linguistic factors affecting attrition and maintenance. The trends suggested by the current study should be further tested, especially in the new immigrant communities that have sprung up recently and that differ markedly from the refugee communities. At the same time, a different kind of language contact situation, especially with global English, has to be considered in the study of colloquial speech of younger generations in present-day Estonia.

References

Altenberg, E. 1991. Assessing first language vulnerability to attrition. In. *First Language Attrition*, H. W. Seliger & R. Vago (eds.), 189–206. Cambridge: CUP.

Bavin, E. L. 1989. Some lexical and morphological changes in Warlpiri. In *Investigating Obsolescence: Studies in Language Contraction and Death,* N. C. Dorian (ed.), 267–286. Cambridge: CUP.

Ben-Rafael, M. 2002. Language contact and attrition: The spoken French of Israeli Francophones.In *First Language Attrition: Interdisciplinary Perspectives on Methodological Issues*, M. Schmid, B. Köpke, M. Keijzer & L. Weilemar (eds), 164–187. Amsterdam: John Benjamins.

Boyd, S. & Andersson, P. 1991. Linguistic change among bilingual speakers of American English and Finnish in Sweden: Background and some tentative findings. *International Journal of the Sociology of Language* 90: 13–35.

de Bot, K., Gommand, P. & Rossing, C. 1991. L1 loss in an L2 environment: Dutch immigrants in France. In *First Language Attrition*, H. W. Seliger & R.Vago (eds), 87–98. Cambridge: CUP.

de Bot, K. & Stoessel, S. 2002. Introduction: Language change and social networks. *International Journal of the Sociology of Language* 153: 1–7.

de Hoop, H. 2003. On the interpretation of stressed pronouns. In *Proceedings of the Conference "sub7- Sinn und Bedeutung"* [Arbeitspapier 114]. Konstanz: FB Sprachwissenshaft, Universität Konstanz. <http://ling.uni-konstanz.de/pages/conferences/sub7/>.

Duvallon, O. & Chalvin, A. 2004. La réalisation zéro du pronom sujet de première et de deuxième personne du singulier en Finnois et en Estonien parlés. *Linguistica Uralica* 40 (4): 270–286.

Ehala, M. 2001. Eesti keele baassõnajärjest. (On the basic word order of Estonian). In *Keele kannul (Following the language),* R. Kasik (ed.), 21–41. Tartu: Tartu Ülikool.

Erelt, M. 2003. Structure of Estonian. *Estonian language*, M. Erelt (ed.), 93–129. Tallinn: Estonian Academy Publishers.

Extra, G. & Verhoeven, L. 1999. Processes of language change in a migration context: The case of the Netherlands. In *Bilingualism and Migration,* G. Extra & L.Verhoeven (eds.), 29–60. Berlin: Mouton de Gruyter.

Grüning, A. & Kibrik, A. A. 2005. Modelling referential choice in discourse: A cognitive calculative approach and a neural network approach. In *Anaphora Processing: Linguistic, Cognitive and Computational Modelling,* A. Branco, T. McEnery & R. Mitkov (eds.), 163–198. Amsterdam: John Benjamins.

Haegeman, L. 2000. Adult null subjects in non-pro-drop languages. In *The Acquisition of Syntax*, M.-A. Friedemann & L. Rizzi (eds.), 129–169. Harlow: Longman.

Heine, B. & Kuteva, T. 2005. *Language Contact and Grammatical Change*. Cambridge: CUP.

Hutz, M. 2002. Is there a natural process of decay? A longitudinal study of language attrition. In *First Language Attrition: Interdisciplinary Perspectives on Methodological Issues*, M. Schmid, B. Köpke, M. Keijzer & L. Weilemar (eds.), 189–206. Amsterdam: John Benjamins.

Jarvis, S. 2003. Probing the effects of the L2 on the L1: A case study. In *Effects of the Second Language on the First*, V. Cook (ed.), 81–103. Clevedon: Multilingual Matters.

Kaiser, E. 2003. The Quest for a Referent: A Crosslinguistic Look at Reference Resolution. PhD dissertation, University of Pennsylvania.

Kaiser, E. & Hiietam, K. 2003. A comparison of the referential properties of third person pronouns in Finnish and Estonian. *Nordlyd, Proceedings of the Workshop on Generative Approaches to Finnic Languages* 31(4): 654–667.

Kaiser, E. 2004. The role of contrast in pronoun interpretation: A look at the referential properties of the Estonian pronoun tema. Paper presented at the 20th Scandinavian Conference of Linguistics, Helsinki.

Kaiser, E. & Trueswell, J. 2004. The referential properties of Dutch pronouns and demonstratives: Is salience enough? In *Proceedings of the Conference „sub8- Sinn und Bedeutung"* [Arbeitspapier Nr.177], C. Meier & M. Weisgerber (eds), 137–149. Konstanz: FB Sprachwissenschaft, Universität Konstanz.

Keevallik, L. 2003. Colloquial Estonian. In *Estonian Language*, M. Erelt (ed.), 343–374. Tallinn: Estonian Academy Publishers.

King, R. 1989. On the social meaning of linguistic variability in language death situations: Variation in Newfoundland French. In *Investigating obsolescence: Studies in language contraction and death,* N. C. Dorian (ed.), 139–148. Cambridge: CUP.

Klaas, B. 2002. Estonians and Estonian language in Southern Sweden and Lithuania. In *Languages in Development*, H. Metslang & M. Rannut (eds.), 37–48. Munich: Lincom.

Köpke, B. & Schmid, M. 2002. Language attrition: The next phase. In *First Language Attrition: Interdisciplinary Perspectives on Methodological Issues*, M. Schmid, B. Köpke, M. Keijzer & L. Weilemar (eds.), 1–43. Amsterdam: John Benjamins.

Labov, W. To appear. Quantitative reasoning in linguistics. <http://www.ling.upenn.edu/~wlabov/Papers/QRL.pdf >.

Lainio, J. & Wande, E. 1994. The Pronoun *minä* 'I' in urban Sweden Finnish. In *The Sociolinguistics of Urbanization: The Case of the Nordic Countries,* B. Nordberg (ed.), 171–202. Berlin: Walter de Gruyter.

Lehiste, I. & Kitching, J. 1998. Sihtise käänete kasutamisest väliseestlaste poolt. In *Väliseestlaste keelest (On the Language of Expatriate Estonians),* L. Lindström (ed.), 67–75. Tartu: Tartu Ülikool.

Lindseth, M. 1995. Null-subject Properties of Slavic Languages: with Special Reference to Russian, Czech and Sorbian. PhDdissertation, Indiana University, Bloomington.

Lindström, L. (ed.). 1998. *Väliseestlaste keelest. (On the Language of Expatriate Estonians).* Tartu: Tartu Ülikool.

Lindström, L. 2001. Verb-initial clauses in narrative. In *Estonian Typological Studies* V. [Publications of the Department of Estonian of the University of Tartu 18], M. Erelt (ed.), 138–168. Tartu: Tartu University Press.

Lindström, L. 2005. *Finiitverbi asend lauses, sõnajärg ja seda mõjutavad tegurid suulises eesti keeles. (The Position of a Finite Verb in a Clause: Word Order and the Factors Affecting it in Spoken Estonian)*. Tartu: Tartu Ülikool.

Maandi, K. 1989. Estonian among immigrants in Sweden. In *Investigating Obsolescence: Studies in Language Contraction and Death,* N. C. Dorian (ed.), 227–342. Cambridge: CUP.

Maher, J. 1991. A crosslinguistic study of language contact and language attrition. In *First Language Attrition*, H. W. Seliger & R. Vago (eds), 67–86. Cambridge: CUP.

Nemvalts, P. 1998. Kas väliseesti keeles on märgata süntaktilist omapära? (Are there notable syntactic characteristics of expatriate Estonian?). In *Väliseestlaste keelest (On the Language of Expatriate Estonians),* L. Lindström (ed.), 55–66. Tartu: Tartu Ülikool.

Oksaar, E. 1972. Spoken Estonian in Sweden and in the USA: An analysis of bilingual behaviour. In *Studies for Einar Haugen,* E. S. Firchow, K. Grimstad, N. Hasselmo & W. A. O'Neil (eds), 437–449. The Hague: Mouton.

Pajusalu, R. 1996. Pronoun systems of common Estonian and Estonian dialects in a contrastive perspective. In *Estonian Typological Studies* I, M. Erelt (ed.), 145–164. Tartu: Tartu University Press.

Pajusalu, R. 1997. Eesti pronoomeneid I: Ühiskeele see, too ja tema/ta. (Estonian pronominals I). *Keel ja Kirjandus*, 24–30 and 106–107.

Pajusalu, R. 2005. Anaphoric pronouns in Spoken Estonian: Crossing the paradigms. In *Minimal Reference: The Use of Pronouns in Finnish and Estonian Discourse,* R. Laury (ed.), 135–162. Helsinki: Finnish Literature Society.

Paolillo, J. 2002. *Analyzing Linguistic Variation. Statistical Models and Methods.* Stanford CA: CSLI.

Pool, R. 1999. About the use of different forms of the first and second person singular personal pronouns in Estonian cases. In *Estonian Typological Studies* III, M. Erelt (ed.), 158–184. Tartu: Tartu University Press.

Raag, R. 1982. *Lexical Characteristics of Swedish Estonian*. Uppsala: Acta Universitatis Upsaliensis.

Raag, R. 1983. *Estniskan i Sverige* [FUSKIS/FIDUS 6]. Uppsala: Uppsala Universitetet.

Raag, R. 1985. The direct object in Swedish Estonian. *Eesti Teadusliku Seltsi Rootsis aastaraamat/Annales Societatis Litterarum Estonicae in Svecia IX* 1980–1984: 201–212.

Raag, R. 1991. Linguistic tendencies in the Estonian language in Sweden. *Linguistica Uralica* 27: 23–32.

Raun, E. 2004, June. Estonians in exile: Continuity and challenge. Paper presented at the 19th conference of the Association for the Advancement of Baltic Studies: Dynamics of Integration and Identity, University of Toronto.

Roos, A. 1980. *Morfologiska tendenser vid språklig interferens med estniska som bas. Acta Universitetis Upsaliensis: Studia Uralica et Altaica Upsaliensia 12*. Uppsala: Uppsala Universitetet.

Sankoff, G. 2002. Linguistic outcomes of language contact. In *Handbook of Sociolinguistics*, J. K. Chambers, P. Trudgill & N. Schilling-Estes (eds), 638–668. Oxford: Basil Blackwell.

Satterfield, T. 2003. Economy of interpretation: Patterns of pronoun selection in transitional bilinguals. In *Effects of the Second Language on the First*, V. Cook (ed.), 214–233. Clevedon: Multilingual Matters.

Seliger, H. W. & Vago, R. 1991. The study of first language attrition: an overview. In *First Language Attrition*, H. W. Seliger & R. Vago (eds.), 3–16. Cambridge: CUP.

Sharwood Smith, M. 1989. Crosslinguistic influence in language loss. In *Bilingualism across the Lifespan: Aspects of Acquisition, Maturity and Loss*, K. Hyltenstam & L. K. Obler (eds), 185–201. Cambridge: CUP.

Siewierska, A. 2003. Reduced pronominals and argument prominence. In *Nominals: Inside and Out*, M. Butt & T. Holloway King (eds.), 119–150. Stanford CA: CLSI.

Toribio, A. J. 2001. On Spanish language decline. In *Proceedings of the Annual Boston University Conference on Language Development,* A. H.-J. Do, L. Domínguez & A. Johansen (eds.), 768–779. Somerville MA: Cascadilla Press.

Vainikka, A. & Levy, Y. 1999. Empty subjects in Finnish and Hebrew. *Natural Language and Linguistic Theory* 17(3): 613–671.

Viitso, T.-R. 2003. Structure of the Estonian language: Phonology, morphology and word formation. In *Estonian language*, M. Erelt (ed.), 9–92. Tallinn: Estonian Academy Publishers.

Yağmur, K. 2002. Issues in finding the appropriate methodology in language attrition research. In *First Language Attrition: Interdisciplinary Perspectives on Methodological Issues*, M. Schmid, B. Köpke, M. Keijzer & L. Weilemar (eds), 130–163. Amsterdam: John Benjamins.

Young, R., & Bayley, R. 1996. VARBRUL analysis for second language acquisition research. In *Second Language Acquisition and Linguistic Variation*, R. Bayley & D. Preston (eds), 253–306. Amsterdam: John Benjamins.

Appendix 1

Estonian personal pronoun system

Case	Form	1pSg	2pSg	3pSg	1pPl	2pPl	3pPl
Nominative	**Long**	**mina**	**sina**	**tema**	**meie**	**teie**	**nemad**
	Short	ma	sa	ta	me	te	nad
Genitive	**Long**	**minu**	**sinu**	**tema**	**meie**	**teie**	**nende**
	Short	mu	su	ta	me	te	
Partitive	**Long**	**mind**	**sind**	**teda**	**meid**	**teid**	**neid**
Illative	**Long**	**minusse**	**sinusse**	**temasse**	**meisse**	**teisse**	**nendesse**
	Short	musse	susse	tasse			neisse
Inessive	**Long**	**minus**	**sinus**	**temas**	**meis**	**teis**	**nendes**
	Short	mus	sus	tas			neis
Elative	**Long**	**minust**	**sinust**	**temast**	**meist**	**teist**	**nendest**
	Short	must	sust	tast			neist
Allative	**Long**	**minule**	**sinule**	**temale**	**meile**	**teile**	**nendele**
	Short	mulle	sulle	talle			neile
Adessive	**Long**	**minul**	**sinul**	**temal**	**meil**	**teil**	**nendel**
	Short	mul	sul	tal			neil
Ablative	**Long**	**minult**	**sinult**	**temalt**	**meilt**	**teilt**	**neilt**
	Short	mult	sult	talt			neilt
Translative	**Long**	**minuks**	**sinuks**	**temaks**	**meieks**	**teieks**	**nendeks**
					meiks	teiks	neiks
Essive, Terminative, Abessive			**long stems only**				
Comitative	**Long**	**minuga**	**sinuga**	**temaga**	**meiega**	**teiega**	**nendega**
	Short	muga	suga	taga			

Appendix 2

Speaker information

Overview of the participants' demographic and immigration data

Name*	Group	Gender	Age	Years in the US	Age of arrival
Henri	1Older	M	70	49	21
Anna	1Older	F	72	52	20
Ella	1Older	F	75	44	31
Juuli	1Older	F	77	50	27
Oskar	1Older	M	70	50+	20+
Linda	1Older	F	69	50	19
Helvi	1Older	F	77	54	23
Agnes	1Older	F	68	52	16
Georg	1Older	M	74	52	22
Vello	1Older	M	72	53	19
Viivi**	1Older	F	70	52	18
Sander	2Younger	M	60	52	8
Riina	2Younger	F	55	52	3
Arno	2Younger	M	65	53	12
Marta	2Younger	F	67	52	15
Ulla	2Younger	F	65	50	12
Juhan	2Younger	M	67	53	14
Diana	3New	F	37	6	31
Meelis**	3New	M	18	4	14
Katrin	3New	F	26	4	23
Toomas	3New	M	30+	4	22
Liina	3New	F	24	3	20
Timo	3New	M	33	8	25

* Pseudonyms are used

** Provided too little speech to be included in the quantitative analysis

Appendix 3

Preference for long pronouns and overt subject total sample, (values closer to 1 show higher preference), individual speakers. (P – personal pronoun); p = .001

Group	Name	Long nonprominent P	Long nonprominent Genitive P	Overt subject P
1 Older	Helvi	0.183	0.335	0.418
2 younger	Riina	0.185	0.171	0.572
1 Older	Valter	0.348	0.192	0.654
1 Older	Agnes	0.353	0.717	0.679
3 new	Diana	0.393	0.552	0.475
2 younger	Juhan	0.397	0.263	0.632
1 Older	Linda	0.437	0.763	0.608
3 new	Timo	0.441	0.064	0.394
3 new	Katrin	0.448	0.596	0.241
1 Older	Juuli	0.467	0.343	0.411
3 new	Liina	0.493	0.701	0.351
2 younger	Ulla	0.507	0.475	0.647
3 new	Toomas	0.515	1	0.274
2 younger	Sander	0.541	0.343	0.755
2 younger	Arno	0.600	0.611	0.260
1 Older	Georg	0.633	0.298	0.735
1 Older	Oskar	0.639	0.819	0.432
2 younger	Marta	0.732	0.788	0.55
1 Older	Anna	0.817	0.666	0.441
1 Older	Henri	0.832	0.760	0.291
1 Older	Ella	0.846	0.632	0.227

Turkish in the Netherlands

Development of a new variety?

A. Seza Doğruöz and Ad Backus
Tilburg University

This paper is about Dutch influence on the variety of Turkish spoken by immigrants in the Netherlands. The community is under constant pressure to shift to Dutch, but maintenance figures are nevertheless very high. The result is a contact situation in which the entire community is bilingual; everyday interaction features much codeswitching and the minority language Turkish undergoes contact-induced changes in both lexicon and grammar. Often, these changes are direct borrowings from Dutch. In this paper, we will be searching for evidence of structural change in the Turkish spoken in the Netherlands (NL-Turkish) by second generation immigrants, and see whether these changes can be traced back to Dutch influence.

1. Introduction

This paper is about Dutch influence on the variety of Turkish spoken by immigrants in the Netherlands. The community is under constant pressure to shift to Dutch, but maintenance figures are nevertheless very high. The result is a contact situation in which nearly the entire community is bilingual, everyday interaction features much codeswitching, and the minority language, Turkish undergoes contact-induced changes in both lexicon and grammar. Often, these changes are direct borrowings from Dutch. In fact, there is a common perception among people in Turkey that the Turkish spoken by the immigrant communities is not the same Turkish anymore, and that a new variety has developed in the diaspora.

As long as the contact situation is intense enough and lasts long enough, anything can be borrowed especially if there is a dominance asymmetry between the groups in contact (Thomason 2001). In order to formalize the relationship between social factors and what tends to get borrowed in a contact situation, Thomason & Kaufman (1988) suggest a borrowing scale. According to this scale, borrowing is confined to lexical items in the initial stages of contact. In bilingual

speech, a diagnostic of this may be a preponderance of insertional codeswitching, in which content words from another language are inserted into a base language structure. The inserted content words in such situations are often similar to the kinds of words that get borrowed over time, e.g. words related to the majority culture in immigrant settings. Such lexical influence was found to be predominant in the Turkish spoken by first generation, Turkish-dominant immigrants in the Netherlands (Backus 1996, Boeschoten 1990).

As contact becomes more intense, structural elements (e.g. case marking and word order patterns) are also borrowed from the other language. Weinreich (1953: 30) refers to these borrowings as "interference" since the grammatical relations of one language (e.g. Dutch) are copied to the other language (e.g. Turkish). This interference is often recognized as a deviation from the norm by the monolingual speakers of the same language, who do not encounter the conventions of the contact language.

In this paper, we will be searching for evidence of structural change in the Turkish spoken in the Netherlands (NL-Turkish) by second generation immigrants, and see whether these changes can be traced back to Dutch influence. Candidates for such structures are those that show differences with Turkish as spoken in Turkey (TR-Turkish). We wish to make a cautionary note from the outset. Deviant structures in contact varieties (such as NL-Turkish), as detected by native speaker judges, are often attributed to the contact language (Dutch, in our case) without question. However, not much is known about the reliability of such judgments. In other words, we usually do not know for sure whether these deviant structures really do not exist in the spoken language of the non-contact variety (i.e. in TR-Turkish, in our case). By analyzing TR-Turkish spoken data, we try to overcome this problem.

The organization of the paper is as follows: in Section 2, we will delimit our domain of inquiry, and discuss our descriptive framework, a combination of Construction Grammar (Goldberg 2005) and the Code Copying Model (Johanson 2002). In Section 3, we will give some background information on the Turkish community in the Netherlands and describe our data and method of analysis. Section 4 will present the research questions, which will be answered in Sections 5 and 6. In Section 7, we will go a little deeper into one particular deviant structure. Finally, Section 8 discusses the results in the light of the question whether these changes are enough justification for saying that the immigrant communities have developed a new variety of the language.

2. What kinds of structures?

A language is made up of a large number of structures. While words make up the lexicon and structures make up the syntax, there are also numerous constructions, which are somehow in between lexicon and syntax, as they exhibit properties of both. We define constructions as conventional units consisting of more than one morpheme irrespective of whether the whole unit is lexically fixed or whether it is partially open (Goldberg 2005, Doğruöz & Backus 2009). Depending on the fixed and open slots they have, constructions occur at various levels, irrespective of traditional notions of constituency. In the following example, different types of constructions can be identified. Basically, every unit that is perceived as conventional and has at least one fixed aspect (form and/or meaning) which qualifies as a construction. As can be seen in Example (1), many constructions contain a mixture of fixed elements and open slots.

(1) *It rains a lot in Holland.*
   a. It rains
   b. It rains ADV PP
   c. It rains a lot PP.
   d. A lot
   e. In N
   f. It V PP
   g. In Holland

There are many studies about lexical and syntactic changes in the language contact literature, but little attention has been paid to this level of language. In this paper, we intend to close this gap by looking specifically at Dutch influence on Turkish constructions. There are signs that Dutch structural influence is mainly found at this level, rather than at that of abstract syntactic structure (Doğruöz & Backus 2007). How this relates to the observation that Turks in the Netherlands seem to be using a new variety of Turkish will be discussed in Section 8.

Among the multitude of terms that are in use for influence from one language on another (e.g. borrowing, calquing, transfer, interference, replication, etc.), we use the term copying proposed by Johanson (2002) since it is more neutral in comparison to the other existing terms. At least two parties are normally involved in a copying procedure: the socially dominated language that serves as the recipient language (NL-Turkish, in this case) and the socially dominating language from which some structures are copied (Dutch, in this case).

Johanson (2002) distinguishes two major types of copying: global and selective copying. Global copying is when overt material from the copied language (e.g. words) is used in the copying language. This kind of copying can be seen in

Example (2), where a lexical item from Dutch is used in a Turkish sentence. This kind of copying is traditionally called **insertional code-switching**.

(2) *Tilburg-ta üç tane BIOSCOOP var.*
Tilburg-LOC[1] three number CINEMA exist.PRES.
'There are three CINEMAS in Tilburg'

As the name implies, in selective copying only some properties of a unit from the other language are taken over. Johanson (2002) distinguishes four types of properties: material, semantic, frequential and combinational.

Material copying involves copying of phonetic properties. If aspects of the meaning of a foreign word are copied onto the semantics of its translation equivalent, this is defined as semantic copying. Sometimes, it is only the frequency of a pattern that is increased or decreased, under the influence of the high or low frequency of its equivalent in the other language. Finally, in combinational copying, co-occurrence patterns are copied from one language to the other. This type is the most important one for our purposes, as it refers to the copying of combination patterns, i.e. conventional combinations of certain morphemes, i.e. what we have referred to as **constructions**. Examples of this kind may include word-order patterns, ways of combining clauses, possessive marking, and so forth. Example (3) illustrates the copying of an agreement pattern from Dutch into NL-Turkish.

(3) NL-Turkish: *Onlar ev- hanım-ları.*
They house-wife- POSS.3PL.
'They are housewives'.
NL: *Ze zijn huis-vrouw-en.*
They are house-wife-PL.
'They are housewives'.
TR-Turkish: *Onlar ev- hanım-ı.*
They house-wife-POSS.3SG.
'They are housewives'.

In Dutch, the subject agrees with the copula in equational copula constructions and this agreement is marked on the copula. In TR-Turkish, on the other hand, there is no number agreement in such constructions. Example (3) shows that NL-Turkish has copied the agreement requirement by marking the complement with the plural marker. Note that there is no overt copula in such sentences. Examples

1. Abbreviations: ABL: ablative, ACC: accusative, ADJ: adjective, COM: comitative, DAT: dative, GEN: genitive, LOC: locative, N: Noun, NEG: negation, OPT: optative, PAST: past, PASS: passive, PL: plural, POSS: possessive, PRES: present, PROG: progressive, QP: question particle. SG: singular.

like this are interpreted as 'unconventional' by TR-Turkish speakers. In this paper, we will be interested in these kinds of combinational copying patterns.

3. Methodology

The Turkish community in the Netherlands was formed as a result of labor migration from Turkey in the 1960's, especially from eastern regions of Turkey. Mostly men took up jobs in several European countries, leaving their families behind. Although immigration was intended to be temporary at first, it became permanent after the family reunifications of the 1970's. Today, the Turkish community is one of the largest immigrant communities (372.714) in the Netherlands (CBS, 2008).

Turks have always maintained close ties with Turkey and Turkish. According to the summary of the relevant literature in Backus (2004), the main factors which assisted in this high level of maintenance of Turkish in the Netherlands are: regular summer-long vacations in Turkey, easy access to Turkish media (TV, newspapers, Internet etc.), and especially the tendency to marry people from Turkey.

For this study, we analyzed three informal conversations in NL-Turkish (11.749 words) and TR-Turkish (12.747 words). The conversations took place between the first author and an NL-Turkish and/or a TR-Turkish informant. These conversations are part of NL-Turkish and TR-Turkish spoken corpora collected by the first author.

The NL-Turkish informants were second generation immigrants, whose parents came from Kırşehir, a small town in Central Anatolia in Turkey. They were all university students between the ages of 18–30. Similarly, TR-Turkish informants were between the ages of 18–30 and were living in Kırşehir, the very place where the parents of the NL-Turkish informants came from. All the TR-Turkish informants were at least high school graduates and were employed in an oil company at the time of the recordings.

4. Research questions

Our starting point was the quest for unconventional constructions in the NL-Turkish data. In other words, if a construction sounds (structurally) unconventional (different) to a TR speaker, it is potentially a structural copy from Dutch. In order to establish unconventionality in NL-Turkish, first we identified all the unconventional NL-Turkish constructions that sounded (structurally) different to the first author, who is a native speaker of Turkish. Second, we consulted five TR-Turkish judges. We showed them the unconventional constructions (in a context) and asked

them to identify anything that sounded different. When they referred to a construction as different, we asked them how they would say the same construction alternatively. Similarly, a group of Dutch monolingual speakers (five) were consulted for the Dutch translations of unconventional NL-Turkish utterances into Dutch. All judges had backgrounds similar to the informants. In addition to unconventionality in NL-Turkish, judges also evaluated non-contact control data (TR-Turkish) for unconventionality and arrived at the following research questions:

1. What types of unconventional constructions are there in NL-Turkish and, if there are any, in TR-Turkish?
2. How often do the unconventional constructions occur?
3. Is there any evidence that NL-Turkish unconventional constructions are copied from Dutch?

The answer to question (3) benefits from comparative evidence from TR-Turkish data. Specifically, if an unconventional construction in the NL-Turkish data turns out to also exist in TR-Turkish, it would make less sense to attribute the unconventionality to Dutch influence.

5. Results I: Unconventionality in NL-Turkish

We first searched for unconventional constructions in our NL-Turkish data, i.e. for constructions that would sound 'unconventional' to a TR-Turkish speaker.

As a first step in the analysis, we categorized these constructions along traditional lines, and found unconventional instances in the following grammatical categories: fixed expressions, noun phrases, verb phrases, adpositional phrases, word order and redundant subject pronouns. However, we will later illustrate that focusing on these traditional categories ignores the subtle differences among the constructions that are influenced by Dutch. Table 1 indicates the frequencies with which unconventional instances were found in these categories for each NL-Turkish speaker (Nahit, Orhun and İsa).

Table 1. Frequency of unconventional NL-Turkish constructions

Grammatical categories	Nahit	Orhun	İsa	Total
Noun phrase (NP)	23	13	10	46
Verb phrase (VP)	13	16	13	42
Adpositional phrases (AP)	11	2	8	21
Word order	9	6	3	18
Fixed expressions	4	0	1	5
Redundant subject	2	0	1	3

As can be seen in Table 1, unconventionality was most frequently observed in noun phrases and verb phrases, and also quite often in adpositional phrases and word order patterns. Below, we will discuss these categories in detail and provide examples.

In Example (4), the adpositional phrase seems to be copied from Dutch; in any case TR-Turkish uses a different phrase in this context.

(4) NL-Turkish: [*hiç fark-ı* *yok* [*İngiliz-le*]]
No difference-POSS.3SG exist.not English-COM.
'There is no difference from English.'
NL: *helemaal* [*geen verschil* [*met Engels-en*[2]]].
Absolutely no difference with English-PL.
'There is no difference from English.'
TR-Turkish: [*hiç fark-ı* *yok* [*İngiliz-den*]
No difference-POSS.3SG exist.not English-ABL.
'There is no difference from English.'

If we look at the example more closely we will notice that the unconventional prepositional phrase is in fact part of a larger construction, both in Turkish and in Dutch. In Turkish, this construction is [*N-den farkı yok*] 'N-ABL. difference exist.not' and it is [*geen verschil met N*] 'no difference with N' in Dutch. Due to influence from Dutch, NL-Turkish speaker uses the comitative *(-le)* instead of the ablative (*-den*). As a result of this change, TR-Turkish speakers identified the NL-Turkish use as unconventional. This sort of adpositional phrase copying has also been observed elsewhere (for example Polinsky 1995 on English-influenced Lithuanian in the USA).

Example (5) illustrates unconventionality in a verb phrase with the [*speak*[3] name.of.language] construction. In the contrastive context of the example, TR-Turkish would mark the direct object (*Türkçe* 'Turkish') with an accusative case marker (*Türkçe-yi* 'Turkish-ACC'); Dutch, on the other hand, doesn't have an accusative case marker. Therefore, the omission of accusative marking in NL-Turkish seems to be copied from Dutch. However, it is too early to attribute this loss only to Dutch influence before finding sufficient evidence.

In Example (5), the speaker had been explaining that the Turkish kids in the Netherlands spoke Dutch well. Therefore, the question about their ability to speak Turkish invokes a comparison with their Dutch.

---

2. In this context, Engels-en "English-PL." refers to the English people in general.

3. Italics indicate the fixed part of the construction.

(5) NL-Turkish: *Türkçe iyi konuş-uyor-lar mı?*
Turkish good speak-PROG-3PL QP.
'Do they speak Turkish well?'
NL: *Sprek-en ze goed Turks?*
Speak-3PL. they good Turkish.
'Do they speak Turkish well?'
TR-Turkish: *Türkçe-yi iyi konuş-uyor-lar mı?*
Turkish-ACC good speak-PROG-3PL QP
'Do they speak Turkish well?'

Example (6) illustrates an NP construction that is perceived as unconventional by TR-Turkish speakers since the Dutch plural-marked construction [*een paar N-PL*] 'a couple N-PL.' seems to induce the NL-Turkish speaker to replace the singular-marked construction [*a couple* N] with the one in which the N is plural.

(6) NL-Turkish: [*Birkaç konser-ler*] *ver- di- ler.*
a.couple concert-PL. give-PAST-3PL.
'They gave a couple of concerts.'
NL: *Ze gav-en* [*een paar concert-en*].
They give.PAST-3PL a couple concert-PL.
'They gave a couple of concerts.'
TR-Turkish: [*Birkaç konser*] *ver-di-ler.*
a.couple consert-PL give-PAST-3PL.
'They gave a couple of concerts.'

Examples (7–9) illustrate the unconventionality in fixed expressions, word order and subject pronoun use. We will not discuss these categories in detail but illustrate them with examples from the data.

Example (7) focuses on the fixed expression of 'I don't know'.[4] TR-Turkish speakers perceive the subject pronoun in this fixed expression as redundant since the NL-Turkish speaker used it in a context which did not involve a contrast or emphasis NL-Turkish speaker used this construction as a conversational filler, and in this context the fixed expression in TR-Turkish is without the subject pronoun. Since the equivalent of this filler construction in Dutch makes use of a subject pronoun, the NL-Turkish speaker also used a subject pronoun in Turkish.

4. If the same expression was used literally, the Turkish translation would have been 'bil-mi-yor-um' (know-NEG-PROG.-1SG) and still would not require the use of a subject pronoun, unless it has a contrastive meaning.

(7) NL-Turkish: *Ben ne bil -e-yim.*
I what know-OPT-1SG.
'I don't know'.
NL: *Ik weet het niet.*
I know that not.
'I don't know'.
TR-Turkish: *Ne bil- e- yim.*
What know-OPT-1SG.
'I don't know'.

Example (8) illustrates the unconventionality in NL-Turkish word order. In TR-Turkish reported speech clauses with the verb *demek* 'say' have the word order of [S finite clause V_DEMEK]. Due to Dutch influence, the word order in the reported speech clause becomes [S V_DEMEK finite clause] in NL-Turkish.

(8) NL-Turkish: *Di-yor-lar bu mesleğ-i sev-me-di-m.*
Say-PROG-3PL. this job-ACC like-not-PAST-1SG.
'They say I did not like this job'.
NL: *Ze zeggen dat ik deze baan niet leuk vond.*
They say.3PL that I this job not good find-PAST
'They say I did not like this job'.
TR-Turkish: *Bu mesleğ-i sev-me-di-m di-yor-lar.*
This job-ACC like-not-PAST-1SG say-PROG-3PL.
'They say I did not like this job'.

Example (9) illustrates the use of a redundant subject pronoun in NL-Turkish. When the topic is maintained in TR-Turkish, there is no need to use a subject pronoun. In the given context, the NL-Turkish speaker was answering the interviewer's question 'Did they come here?', which includes a subject pronoun. As an answer to this question, there is no need to use a subject pronoun (neither the directional adverbial) in TR-Turkish since the topic is maintained. However, NL-Turkish speaker keeps the subject pronoun (probably due to Dutch influence). This, as a result, is perceived as unconventional by TR-Turkish speakers.

(9) Interviewer: *Onlar bura-ya gel-di mi?*
They here-DAT come-past QP?
'Did they come here?'
NL-Turkish: *Onlar bura-ya gel- me- di.*
They here-DAT come-NEG-PAST.
'They did not come here'.

NL: *Ze kwamen hier niet.*
They come-PAST here not.
'They did not come here'.
TR-Turkish: *Gel- me- di.*
Come-NEG-PAST.
'They did not come here'.

Based on the brief analyses we have made, the following tentative generalization is possible: there are some unconventional constructions in NL-Turkish, and some of these bear a striking resemblance to equivalent Dutch constructions.

6. Results II: Unconventionality in TR-Turkish

As explained earlier, in order to be able to attribute any unconventionality in NL-Turkish to Dutch, we should be certain that these constructions do not exist in TR-Turkish. Therefore, we also analyzed the TR-Turkish data.

Contrary to what one might expect, we also came across some unconventionality in the TR-Turkish data. This is even more surprising if one considers that it is convenient for us to attribute the unconventionality in NL-Turkish to Dutch influence.

We classified the unconventionality in the TR-Turkish data in a way similar to the classification of the NL-Turkish data. In the TR-Turkish, we found unconventionality in the following categories: noun phrases, verb phrases, fixed expressions and word order. As can be seen in Table 2, there are fewer categories of unconventionality and the frequency of occurrence within these categories is lower compared to NL-Turkish. Since the types of unconventionality are similar to the ones observed in NL-Turkish, we will not provide examples of each type of unconventionality in TR-Turkish.

Table 2. Frequency of unconventional constructions in TR-Turkish

Grammatical categories	Diren[5]	Mahsun	Yetkin	Total
Noun phrase	3	14	3	20
Verb phrase	1	8	2	11
Fixed expressions	0	1	0	1
Word order	2	0	0	2

5. Diren, Mahsun and Yetkin are the names of the TR-Turkish informants.

Unconventionality in TR-Turkish existed mostly in noun phrases and verb phrases. Example (10) illustrates unconventionality in a verb phrase.

(10) TR-Turkish: *Ben Kırşehir yemek-leri bil-ir-im.*
I Kırşehir dish-POSS.3PL know-PRES-1SG.
'I know Kırşehir dishes'.
Original construction: *Ben Kırşehir yemek-leri-ni bil-ir-im.*
I Kırşehir dish-POSS.3PL-ACC know-PRES-1SG.
'I know Kırşehir dishes'.

Apparently, the construction [OBJ-ACC *know*] may sometimes lose its accusative marking in TR-Turkish. When asked, TR-Turkish judges recognized this example as unconventional, but, apparently, these constructions are sometimes used anyway.

Recall that when a similar unconventional construction was used by a NL-Turkish speaker in Example (5), we attributed this to Dutch influence. However, if TR-Turkish speakers use it as well, attributing the unconventionality to Dutch in NL-Turkish becomes questionable. In other words, if we had not known that the informant responsible for this example was a TR-Turkish monolingual, we would probably have attributed the unconventionality to the influence of Dutch, since Dutch does not have accusative marking.

Similarly, in Example (11), the TR-Turkish speaker uses an unconventional construction, a plural NP, that must strike the reader as oddly familiar by now.

(11) TR-Turkish: *Burda* [*birçok çeşit lastik-ler*] *yap ıl- ıyor.*
Here a.lot.of kind tyre-PL. make-PASS-PROG.
'A lot of different types of tyres are made here'.
Original construction: *Burda* [*birçok çeşit lastik*] *yap-ıl-ıyor.*
Here a.lot.of kind tyre make-PASS-PROG.
'A lot of different types of tyres are made here.'

Instead of the original construction [*a.lot.of kind* N], TR-Turkish speaker used the [*a.lot.of kind* N-PL.] construction. This example is similar to the unconventional NL-Turkish construction in Example (6), which was attributed to Dutch influence. Apparently, it is not that straightforward. It seems like some internal change is taking place in TR-Turkish as well. However, we leave this point for a further study.

In order to look further into the similarities and differences between unconventionality in NL-Turkish and in TR-Turkish, we decided to single out one category, the noun phrases.

7. Comparison of unconventional constructions in NL-Turkish and TR-Turkish noun phrases

First, we identified all unconventional noun phrases in the two data sets. The results indicate that NL-Turkish is different than TR-Turkish in terms of the types and the frequencies of the unconventional constructions.

As for frequency (see Table 3), NL-Turkish made use of unconventional constructions significantly more often than TR-Turkish ($t_4 = 5.34$, $p = .006$, $p > .005$). This suggests that contact has some structural effects.

Table 3. The frequency of unconventional NP constructions

	Total number of NP's	Unconventional NP constructions	Ratio
NL-Turkish	889	46	5.0 %
TR-Turkish	1190	20	1.6%

The two sets of data are also different in terms of the qualitative features of these constructions as well. Certain types of unconventional constructions (within the NPs) only existed in NL-Turkish but not in TR-Turkish, cf. Table 4.

Table 4. Types of unconventionality within the NP's

Types of unconventionality (within NP)	NL-Turkish	TR-Turkish
Omission of genitive	2	6
Addition of possessive	2	3
Addition of an indefinite article	6	4
Addition of genitive	1	1
Addition of plural	3	4
Replacement of a lexical item	5	2
Word order change	8	0
Addition of a lexical item	7	0
Omission of possessive	5	0
Unconventional derivational morpheme	7	0
Total	**46**	**20**

The following categories of unconventionality are unique to NL-Turkish.

Addition of a lexical item:

(12) NL-Turkish: *Türk* *insan-lar*
Turk person-PL.
'Turkish people'

NL: *Turkse mensen*
Turkish people
'Turkish people'
TR-Turkish: *Türk-ler*
Turk-PL.
'Turks'

Omission of possessive:

(13) NL-Turkish: *dil.kurs*
language.course
'Language course'
NL: *taalcursus*
tanguage.course
'Language course'
TR-Turkish: *dil kurs-u*
language course-POSS.3SG
'Language course'

Unconventional derivational morpheme:

(14) NL-Turkish: *kimya-sal mühendis-i*
chemistry-ADJ engineer-POSS.3SG
'chemical engineer'
NL: *chemisch ingenieur*
chemical engineer
'chemical Engineer'
TR-Turkish: *kimya mühendis-i*
chemical engineer-POSS.3SG.
'chemical engineer'

Word order change inside the NP is the most frequent type of unconventionality in NL-Turkish (see Table 4). Example (15) illustrates an unconventional NP construction that is influenced by Dutch word order.

(15) NL-Turkish: *strüktür-ü bir organize-nin*
structure-POSS.3SG one organization-GEN.
'structure of a organization'
NL: *struktuur van een organisatie.*
structure of an organization.
'structure of an organization'
TR-Turkish: *bir organize-nin strüktür-ü*
One organization-GEN structure-POSS.3SG.
'structure of an organization'

Looking at the Dutch equivalent, it seems like the Dutch [NP *van* NP] 'NP of NP' construction has been partially translated into NL-Turkish. The order of the two nouns is Dutch-like, though note that the functional elements remain suffixed to their heads: no Dutch-style genitive preposition is developed, nor is the possessive morpheme on the head noun omitted (Dutch does not make use of such a morpheme for this construction. Therefore, Dutch influence could have triggered the omission of the possessive morpheme in NL-Turkish). Since we did not encounter this kind of deviation in the TR-Turkish data, we seem to be on safer ground than with Examples (5) and (6) to attribute this deviation to Dutch influence. Further support comes from similar changes attested in other contact settings (e.g. Leisiö 2000 and Hill & Hill 2004).

8. Discussion

Our starting point in this paper was to investigate whether there are any unconventional constructions in NL-Turkish and if so, whether we would be able to attribute the unconventionality to Dutch influence. As can be seen in Section 5, both questions could be answered positively. However, in order to back up the claim of Dutch influence, we also analyzed TR-Turkish data to make sure that these unconventional constructions existed only in NL-Turkish.

Counter to our expectations, we found that certain unconventional constructions also existed in TR-Turkish, though they occurred less frequently and did not come in as great a variety. But still, some deviations which, at first glance, could be attributed to Dutch influence, were also found in TR-Turkish (see Examples 5 and 10; 6 and 11). This testifies to the importance of 'non-contact' data in contact linguistics, since our knowledge of what goes on in the spoken variety of any language may be less than perfect. As Dabrowska (2004) has pointed out, corpus data are invaluable for analyses of actual use and also to find structures that are supposed to be ruled out by grammar.

It is beyond doubt that Dutch influences Turkish in the contact setting. Earlier work, summarized in Backus (2004), has documented extensive codeswitching and, therefore, lexical borrowing, and at least many incidental cases of structural borrowing. What has not been documented, however, is that whether the immigrant variety has undergone systematic structural changes. To a certain extent, this is no doubt due to the fact that little systematic study of this issue has been done so far. However, our study points to another explanation for this fact. Though there are clearly many instances of unconventional structure in NL-Turkish, these are not systematic in the sense that they apply across the board in entire subsystems of the language, such as, clausal word order, NP-internal word order, or

accusative case marking. Instead, we find unconventional constructions in isolated instantiations of these subsystems, and, very often, these are more or less faithful translations of the equivalent Dutch expressions. Dutch influence, we conclude, is item-based rather than structural in the traditional sense. For this reason, it makes little sense to categorize the unconventional constructions using traditional syntactic categories, like we did initially in Table 1. We hypothesize that sweeping changes in entire subsystems will only occur after such item-based changes have increased in frequency (cf. Rostila 2006). NL-Turkish speakers will only then start to change their abstract representations from the original Turkish structure to the new Dutch-influenced one, on the basis of all these individual expressions that instantiate the new structure (cf. Croft 2000, Croft & Cruse 2004, and Tomasello 2003 for the mechanisms behind such a process). However, NL-Turkish is not yet at this stage. This can be related to the present Dutch-Turkish contact situation. As we have mentioned in Section 3, the Turkish community in the Netherlands has strong ties with Turkish as spoken in Turkey. Although the informants in this study had lived their whole life in the Netherlands, due to these strong ties with Turkey, their Turkish is only influenced by Dutch in isolated expressions (i.e. loan translations) rather than systematically. Finally, it should be pointed out that the unconventional constructions are rarely exact translations. The results are hybrids in nature, carrying both Dutch and Turkish structural characteristics. A challenge for future work is to establish the principles that govern which elements within a construction are susceptible to change, and which are not.

References

Backus, A. 1996. *Two in One. Bilingual Speech of Turkish Immigrants in the Netherlands.* Tilburg: Tilburg University Press.

Backus, A. 2004. Convergence as a mechanism of language change. *Bilingualism: Language and Cognition* 7: 179–181.

Boeschoten, H. E. 1990. *Acquisition of Turkish by Immigrant Children: A Multiple Case Study of Turkish Children in the Netherlands aged 4 to 6*. Wiesbaden: Harrassowitz.

CBS. 2008. Bevolking naar herkomstgroepering en generatie. <www.cbs.nl>.

Croft, W. 2000. *Explaining Language Change. An Evolutionary Approach*. Harlow: Pearson Education.

Croft, W. & Cruse, A. D. 2004. *Cognitive Linguistics.* Cambridge: CUP.

Dabrowska, E. 2004. *Language, Mind and Brain: Some Psychological and Neurological Constraints on Theories of Grammar*. Washington DC: Georgetown University Press.

Doğruöz, A. S. & Backus, A. 2007. Postverbal elements in immigrant Turkish: Evidence of change? *International Journal of Bilingualism* 11: 185–220.

Doğruöz, A. S. & Backus, A. 2009. Innovative constructions in Dutch Turkish: An assessment of on-going contact-induced change. *Bilingualism: Language and Cognition* 12(1): 41–63.

Goldberg, A. 2005. Constructions at work. The nature of generalization in language. Oxford: OUP.
Hill, J. & Hill, K. 2004. Word order type change and the penetration of Spanish de in modern Nahuatl. *Sprachtypologie und Universalienforschung* 57: 1–23.
Johanson, L. 2002. *Structural factors in Turkic language contacts*. Richmond: Curzon Press.
Leisiö, L. 2000. The word order in genitive constructions in a diaspora Russian. *The International Journal of Bilingualism* 3: 301–325.
Polinsky, M. 1995. Cross-linguistic parallels in language loss. *Southwest Journal of Linguistics* 14: 87–123.
Rostila, J. 2006. Storage as a way to grammaticalization. *Constructions* 1: 1–59.
Thomason, S. G. 2001. *Language Contact: An Introduction*. Washington DC: Georgetown University Press.
Thomason, S. G. & Kaufman, T. 1998. *Language Contact, Creolization and Genetic Linguistics*. Berkeley CA: University of California Press.
Tomasello, M. 2003. *Constructing a Language: A Usage-based Theory of Language Acquisition*. London: Harvard University Press.
Weinreich, U. 1953. *Languages in Contact.* New York NY: Linguistic Circle of New York.

The reflection of historical language contact in present-day Dutch and Swedish

Charlotte Gooskens, Renée van Bezooijen
and Sebastian Kürschner
University of Groningen

In the present study we quantitatively examine similarly constructed samples of formal spoken Swedish and Dutch in order to compare the composition of the lexicons. Results showed that Swedish has many more loans than Dutch, namely 44.4% against 27.9%. Within the Swedish loans there is a large compartment of Low German (38.7%), whereas most loans in Dutch have a French origin (63.8%). The differences in terms of the number and distribution of loanwords between the lexical profiles of Swedish and Dutch appear to be stable, as they were attested both in the present study and in previous studies. They can be attributed to differences in the linguistic distances between source and borrowing languages and to differences in the intensity of the contacts.

1. Introduction

It is well known that language contact may result in linguistic changes at all levels: phonetic, phonological, morphological, syntactical, prosodic and lexical, depending, among other things, on the intensity of the contact and the degree of relatedness of the languages involved. In this article we will focus on the lexical level, which is generally assumed to be the level that is most easily influenced (Thomason 2001: 69). We were interested to know to what extent different language contact histories may lead to differences in the composition of the lexicon of present-day languages. We opted for comparing Swedish and Dutch. These two Germanic languages share many stems due to their common origin in Proto-Germanic. So, originally their lexicons were very similar. However, the two languages have diverged considerably, as a consequence of both language internal and language external factors, in particular language contact. The nature of these contacts is well documented. In the course of time, Dutch and Swedish have been in contact with the same languages, particularly Low and High

German, Latin and French, but the intensity and duration of these contacts differed considerably (see Section 2). It is not our intention to throw new light on these historical developments. Our research question is: How are the similarities and differences in language contact in the past reflected in the Dutch and Swedish languages as they are used at present?

The few studies that have assessed the composition of the present-day Dutch and Swedish lexicons in a quantitative way were exclusively based on newspaper texts (see Section 3). Moreover, different methodologies and different types of texts were used, which makes it difficult to compare the results. For our contrastive investigation, we took great care to ensure that the databases for the two languages were constructed in exactly the same manner. Both databases include prepared speeches and spontaneous dialogues that can be characterized as formal. All material originates from meetings held in the European Parliament in the first months of 2000, either in Dutch or in Swedish. In this way, the content and style of the speech material was kept constant. In view of the setting (monologues and dialogues for a large public) and topics (politics, economics, administration), the style can be characterized as formal. Moreover, for both Dutch and Swedish we only looked at the most frequent words, as the analysis of frequent words will be less prone to chance fluctuation than the analysis of infrequent words. Finally, the same procedure of word selection and coding was applied to both languages (see Section 4). All this taken together, we provide a quantitative study based on a large and reliable data set that is suitable for a valid comparison of the two languages involved.

Our research questions can be formulated as follows:

1. What are the proportions of inherited words and loanwords in contemporary Dutch and Swedish?
2. What are the origins of the loanwords in the two languages?
3. Which historical prerequisites, such as language contact situation or linguistic distance, can help to explain the differences between the lexical profiles?

2. Historical background

The histories of loanwords in Dutch and Swedish reveal similarities as well as differences.[1] In the early Middle Ages, both languages borrowed many Latin

1. For the history of loans in Dutch, cf. e.g. Van der Sijs (2005), for Swedish cf. Edlund & Hene (2004).

and some Greek words as part of Christianization. Throughout the Middle Ages, Latin remained influential because of its leading role in the church and in the sciences.

In the late Middle Ages, both the Swedish and Dutch dialects had intensive contacts with Low German. This was the language of the Hanseatic League, which constituted a strong economic power. The Hanseatic merchants, who were located in Northern Germany (mainly Lübeck and Hamburg), built up a trade network covering all of Northern Europe. As a consequence of the closer relationship between West Germanic Dutch and Low German, the dialects which would later constitute the basis of Dutch shared many words with Low German. The dialect contacts in the Hanseatic era may have changed the frequency of use of some native words, but this did not lead to intense borrowing. In contrast, North Germanic Swedish, with fewer parallels in the lexicon, was altered by the Low German influence, which is evident from a large number of loanwords.

Already in the early medieval period the Dutch dialects were strongly influenced by French (Van der Sijs 1996: 134). This can be inferred from the great number of loans that go back to Old French. Examples are *vijg* 'fig' from O.Fr. *figue*, *kussen* 'pillow' from O.Fr. *cussin* or *coissin*, *prei* 'leaks' from O.Fr. *porée*, *aalmoes* 'alms' from O.Fr. *almosne*, *toren* 'tower' from O.Fr. *tur* and *fel* 'fierce' from O.Fr. *fel*. At this point of time, the centre of economic power in the Dutch-speaking dialect area was located in the Southern Flemish part, which was strongly influenced, both economically and politically, by the neighboring French-speaking area. The influence of French on Swedish started much later, in the seventeenth and eighteenth centuries. In both Dutch and Swedish the dominant position of French remained stable until the nineteenth century. The upper classes were bilingual, and used many French words in their native language. As their manner of speaking had prestige, the French words were adopted by the middle and lower classes, and thus got incorporated into the general language (Van der Sijs 1996: 139).

The standardization of the West and North European written languages started in the sixteenth century. With the translation of the Luther bible from High German into Dutch in the first half of the sixteenth century and with a considerable number of Germans living in the Dutch language area, High German gained influence on Dutch. From the time of the reformation and the Thirty Years' War (1618–1648) onwards, a comparable development took place in Swedish. In the nineteenth century, High German influence was at its top, especially in the domains of science, industry and trade. In the twentieth century, and especially after the Second World War, English words started to be adopted in Dutch and Swedish, especially within the domains of industrialization, transport, technology and sports. In fact, English is now almost the sole provider of loanwords.

In contrast to Swedish, the Dutch language has (partly) been shaped by its colonial history, specifically in Indonesia. This influence was not as strong, however, as that of the European languages mentioned above.

In Figure 1 we present a schematized overview of the intensity and duration of the language contacts for Dutch and Swedish. Low and High German are represented by the same line. Low German contacts mainly took place in the Middle Ages (before 1550), whereas High German contacts mainly occurred from early modern times (after 1550). The two contact situations can thus easily be distinguished. The contacts with Low German-speaking Hanseatic merchants from Northern Germany involved the entire population in the neighboring countries. The use of High German, in contrast, was largely restricted to the court and institutions of higher education. Moreover, whereas the contact with Low German was mainly established via the spoken language, the contact with High German took place mainly via the written channel of communication (cf. Braunmüller 2004: 23). This is why we consider the former to have been more pervasive, affecting larger portions of the population than the latter. The Dutch and Swedish contacts with Low and High German were comparable in intensity.

The contact between Dutch and French was more intensive than that between Swedish and French. French was only a court language in Sweden, whereas it was a high prestige neighboring language of the Dutch language area. The duration of intensive contact between Dutch and French was much longer than between Swedish and French, since it started already in the Middle Ages. For Dutch, the contact with French was intensive on the written as well as on the spoken level, and for large parts of the population. In Swedish, by contrast, the contact was

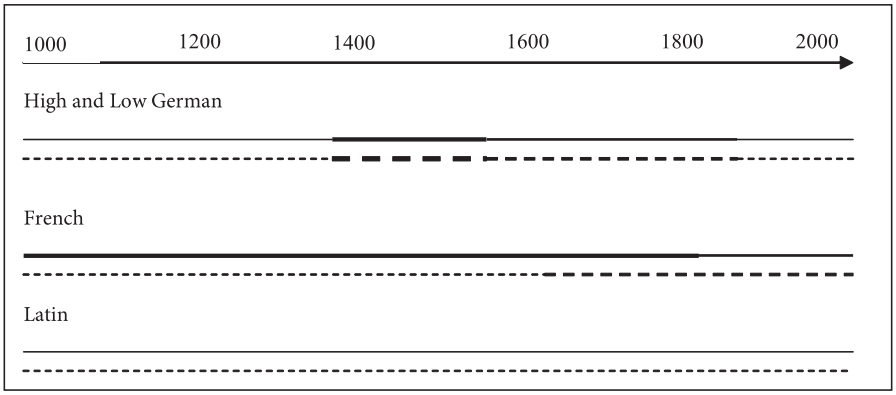

Figure 1. Schematized intensities of language contacts for Dutch (full lines) and Swedish (dotted lines) between 1000 and 2000; the degree of intensity is indicated by boldness

mainly written. The use of spoken French was restricted to a small minority, specifically the influential, highly educated parts of the population.

Both Dutch and Swedish had contact with Latin. In both languages, this contact took place on the written level and was therefore only accessible to a small part of the society for a long period of time. The intensity of this language contact therefore seems comparable for Dutch and Swedish, as indicated in Figure 1.

3. Previous investigations

Van der Sijs (1996: 65) presents a small exploratory study of the first four pages of *NRC Handelsblad* (one of the major daily newspapers in the Netherlands) of April 7, 1994, totaling 11,872 words. After (1) removing personal and geographic names, (2) collapsing conjugated verb forms, and (3) splitting up compounds, she retained 2,144 different lexemes. 69.3% of these were inherited words, dating back to the time when the Germanic languages still formed a unity, and 30.7% were loanwords. Apparently, Van der Sijs is quite impressed by the high number of loans, for she states that "Dutch has received, and is still receiving, loanwords warmly and hospitably" (our translation). The large majority of the loans in the *NRC*-sample have a Latin, French, Italian or Spanish origin (82.0%). High and Low German contributed 6.8% of the loans and English 7.4%. The few remaining loans, totaling 3.7%, are from Greek, Arabic, Hebrew, Turkish and Celtic.

Gellerstam (1973) made an analysis of the 6,000 most frequent word forms found in a Swedish frequency dictionary (Allén 1970). This dictionary is based on one million words from five different Swedish newspapers from 1970. 42.2% of the words in the texts were inherited words, whereas 47.2% consisted of loanwords. The rest were words of unknown origin (1.6%) or words from word classes which Gellerstam excluded from his analysis, i.e. names, numbers etc. (9.1%). Most loans had a High or Low German (51.1%) or a Latin/Greek (43.0%) origin. Only few loans originated from French (2.3%) or English (0.8%) or other languages (2.8%).

A comparison of the data reported for Dutch by Van der Sijs (1996) and for Swedish by Gellerstam (1973) suggests that the Swedish lexicon contains many more loans than Dutch (47.2% versus 30.7%) and that the extent to which different languages have contributed to the lexicons of the two languages differs. Swedish appears to have borrowed more from High and Low German than Dutch has (51.1% versus 6.8% of the loanwords), whereas Dutch appears to have borrowed more from Romance languages than Swedish (82.0% versus less than 45.3% of the loanwords). However, this conclusion can only be tentative, as the two samples of newspaper texts differ in several respects. The Dutch database is small and a mixed sample of frequent and infrequent words, whereas the Swedish database

is large and restricted to high-frequency words. Moreover, the Dutch database comprised lexemes (with compounds broken up into their constituent elements), whereas the Swedish sample is based upon word forms. Moreover, there is a time lapse of more than twenty years between the two samples.

4. Method

4.1 Material

For our investigation we made use of the so-called *Europarl* corpus, which can be downloaded from the internet.[2] This is a parallel corpus that is available for eleven European Community languages, including Swedish and Dutch. Each language is represented by approximately 28 million words. The corpus consists of monologues and dialogues by speakers and chairpersons, collected during meetings in the European Parliament. Both the original speech and the simultaneous translations by the interpreters into the various languages are included. In this way, the same texts can be selected. This is important for our purpose, since we wanted to compare the lexical profiles of Dutch and Swedish keeping subject matter and style constant. For the present study, we selected all the words from the meetings that were held between January 17 and March 17, 2000 for either language, which sums up to roughly one million words per language.

For our investigation we used the 1,500 most frequent words in the one-million-word databases of each of the two languages. The frequency data were gathered for lexemes rather than word forms. This means that the frequencies of, for example, *huis* 'house' and *huizen* 'houses' were added together. In the frequency database, the lexeme is represented by the singular form *huis.* Verbs are represented by the infinitive forms and adjectives by the undeclined form. In the *Europarl* corpus, the Dutch words had already been lemmatized. The Swedish words we lemmatized ourselves by means of The Granska tagger from The Royal Institute of Technology in Sweden.[3] Out of the 1,500 most frequent words we removed all personal names, geographical references and interjections. Compounds constitute a problem, as they may contain words with different etymologies. This is why we split up all transparent compounds into their simple stems. In this way each part could be categorized separately. Examples of such compounds in Dutch are *badkamer* 'bath room' and *vrijdagavond* 'Friday night'. Swedish examples are *fredsprocess* 'peace process' and

2. http://www.statmt.org/europarl/

3. The Granska tagger is available for download online via http://www.csc.kth.se/tcs/humanlang/tools.html (accessed December 11, 2006).

arbetsgrupp 'working group'. In accordance with the procedure adopted by Van der Sijs (1996, see Section 3), compounds with a preposition as the first element were not split up (e.g. Dutch *opbellen* 'call' and Swedish *avsluta* 'finish'). Eventually 1,400 Dutch and 1,418 Swedish lexemes remained for further analysis.

4.2 Coding

Each word was given codes that contained the following information:[4]

1. Inherited word or loanword
2. For loanwords: language from which it has been borrowed directly

For Dutch, the etymological information in (1) and (2) was taken from Van der Veen & Van der Sijs (1997). If the information was not found there, Van der Sijs (1996) was consulted. The Swedish information was found in Wessén (1960) and Hellquist (1980).

5. Results and interpretation

5.1 The proportion of inherited words and loanwords

In Figure 2, the percentages of inherited words and loanwords are shown for Dutch and for Swedish. In Figure 3, the percentages of loanwords in our investigation are compared with the percentages reported in previous investigations. The percentage of loanwords in Dutch (27.9%) is similar to the percentage which Van der Sijs (1996) found in a small newspaper corpus from 1994 (30.7%). Also the percentage of Swedish loanwords (44.4%) is similar to the percentage of loanwords found in the previous investigation by Gellerstam (1973) in a newspaper corpus (47.2%). So for both languages the distribution of inherited words and loanwords is almost identical in two types of formal speech, namely written language in newspapers and monologues and dialogues from the European parliament. Apparently, the composition of the present-day Dutch and Swedish lexicons is very stable in this respect. Both the previous studies and our own study show that Swedish has more loanwords, and consequently fewer inherited words, than Dutch. To gain insight into the nature of this difference, we looked at the quantitative contribution of the source languages in closer detail.

4. The following information was added as well, but not used in the present investigation: original language, year of introduction into the language, word class, pronunciation, word length, cognate/non-cognate.

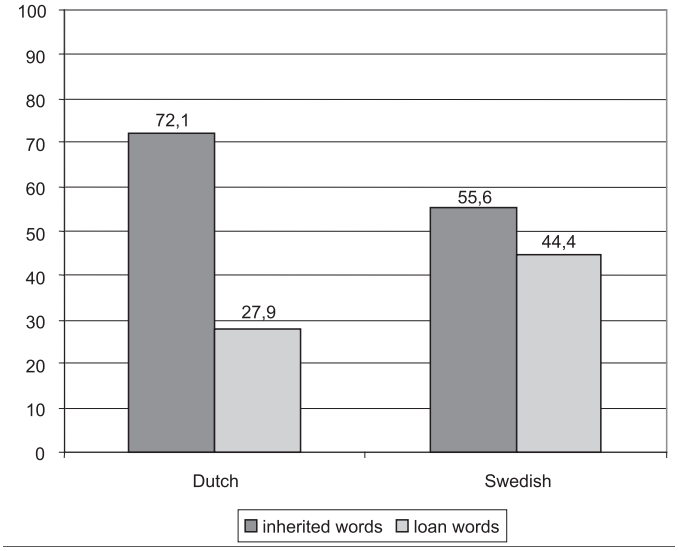

Figure 2. Percentage of inherited words and loanwords in the Dutch and Swedish Europarl corpora

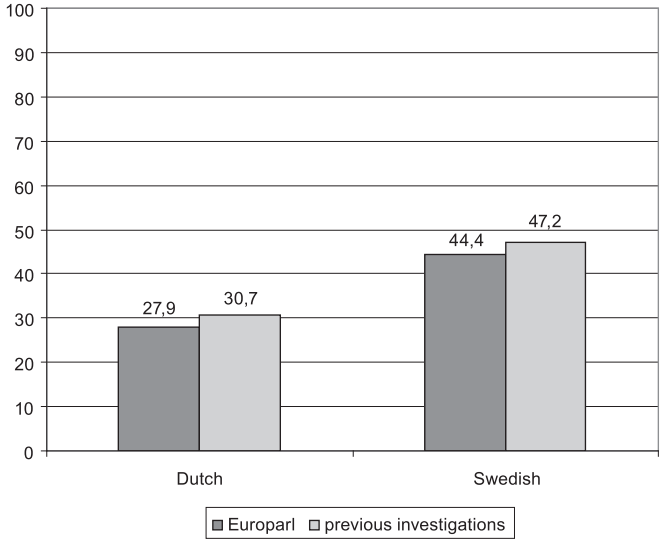

Figure 3. Percentages of loanwords in Dutch and Swedish in the present Europarl corpus and previous investigations (Van der Sijs (1996) for Dutch and Gellerstam (1973) for Swedish)

5.2 The contribution of different languages and the relation with language contact

In Figure 4, the origin of the loanwords in Dutch and Swedish is broken down for source language. Two differences stand out. First, Swedish has a large percentage of Low German loans (38.7%) whereas Dutch has none. Second, Dutch has many more French loanwords (63.8%) than Swedish (14.6%). The other differences are much smaller. For example, Swedish has slightly more Latin and Greek loans than Dutch (differences of 2.6% and 4.3%, respectively). The number of High German loans is higher in Swedish as well (a difference of 5.1%).

Let us at first take a look at the German loans. In Section 2, we reported that both language communities underwent considerable influence of Low German due to the intense contacts with the Hanseatic tradesmen in the Middle Ages. According to our results, however, this parallel contact situation did not result in equal amounts of lexical borrowing. While in the Swedish sample Low German constitutes the largest group of loans (38.7%), Dutch has no Low German words, or at the most one. In fact, of the 36 Dutch words with a German origin, 35

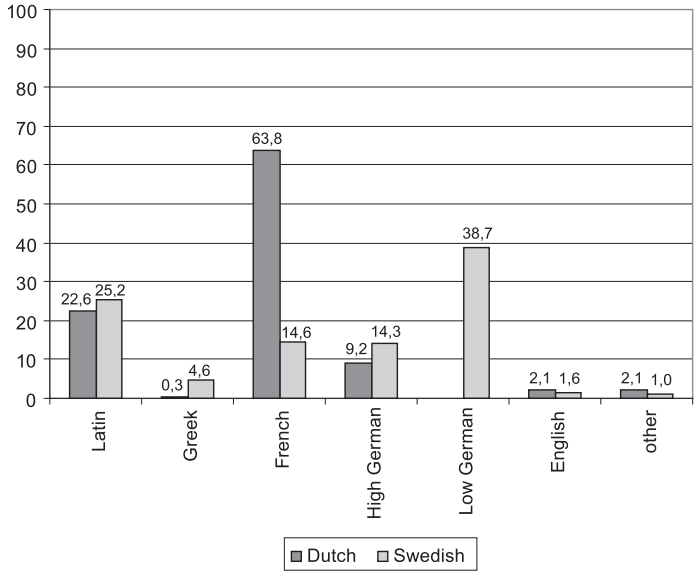

Figure 4. Origin of the loanwords (proportions of the total number of loanwords) in the Europarl corpora

are attributed unambiguously to High German, whereas the precise origin of one word, namely *grens* 'border', is unclear. It could either go back to the Low German *grenize* or to the High German *grenze*.

The difference between the numbers of Low German loanwords in the two languages can be explained when language distances are taken into account. While Middle Swedish, as a North Germanic language, already diverged considerably, structurally as well as lexically, from Middle Low German, Middle Dutch and Middle Low German were part of the same dialect continuum, where mutual intelligibility was highly probable (Goossens 2000). This holds for the grammar as well as for the lexicon. As Dutch and Low German were so similar, there was little room for borrowing. As we pointed out in Section 2, the intense dialect contact is likely to have changed the frequency of single words that were part of both the Low German and Dutch vocabularies. These words are often unidentifiable, though. We are also confronted with the problem that it is difficult to establish the etymology of some words from the available sources of Dutch language history. When a word is found in Dutch as well as in Low German and High German documents from the Middle Ages, it is highly probable that this word is originally a West Germanic word. It cannot be completely excluded, however, that it was introduced to one of the regions only later as a loan. Due to these difficulties, the true origin of many loans from Low German may be concealed in Dutch because we interpret them incorrectly as common West Germanic: "The agreement between Low German and Dutch sometimes makes it difficult to decide whether a word is borrowed or related" (Van der Sijs 1996: 231; our translation).

The contact situation in Sweden was one of structurally related languages as well. Braunmüller (1995) even assumes that semi-communication, i.e. a situation where languages are so alike that their speakers can communicate each using their own language, may have been possible between Scandinavians and Low Germans at this point of time. Nevertheless, the lexical differences between Middle Swedish and Middle Low German were much larger than between Middle Dutch and Middle Low German. This makes it easier to identify Low German loans in modern Swedish.

The high number of Low German loans in our Swedish database can also partly be explained by the loan of derivations: Through prefixation in Low German, the same roots can appear in different lexemes, which were then borrowed into Swedish – like the root *sluta* in *ansluta* 'connect', *avsluta* 'finish', and *besluta* 'decide'. Of the 69 Swedish verb stems of Low German origin in our database, only 55 bear different roots. As Diercks (1993) has shown, affixes of this kind have even become modestly productive in Swedish.

Swedish has fewer High German (14.3%) than Low German loans (39%), but still more than Dutch (9.2%). This difference can be attributed, at least partly, to

the same factor that we mentioned above to explain the differences for Low German, namely the larger linguistic distance between Swedish and Low German than between Dutch and Low German. Moreover, it should be noted that the High German influence, which mainly took place in the seventeenth and nineteenth centuries, was less intense than that of Low German in Swedish.

With respect to French and Latin loans, the differences between Dutch and Swedish cannot be attributed to the same linguistic and historical factors as for the High and Low German loans. The contact with French was more intense in the directly bordering Dutch-speaking area over a much longer period (from the early Middle Ages on) than in the non-bordering Swedish-speaking area, where French words mostly entered in the high-prestige times of French as a court language in the seventeenth and eighteenth centuries. The long-lasting contact situation of Dutch resulted in a considerably larger number of loanwords (63.8%) compared to Swedish (14.6%).

The long-term language contact with French brings about some methodological problems for determining Latin loans in Dutch. In Dutch, many Latin words were not borrowed directly but via French. Direct borrowing from Latin was more frequent in Swedish. As our data pertain to the direct loan-giving language, Latin words which were imported via French were counted as loans from French. For example, a loan like Dutch *civiel* / Swedish *civil* counts as a French loan (of Latin origin) in Dutch, whereas it was categorized as a Latin loan in Swedish. The main cause of the higher number of Latin loans in Swedish is, therefore, the route along which borrowing took place. It should also be noted that in many cases it cannot be established whether a Latin loan was adopted directly from Latin or via French. In our database the Dutch dictionary indicated 'French or (Medieval) Latin' in approximately 5% of the cases. In these cases we categorized the words as French loans, which means that there may have been a slight bias towards French loans in our Dutch database. Many French loans in Dutch can easily be identified to be part of the "Euro-Latin" used in the formal speech of many European languages, and are identical to the corresponding Swedish Latinisms.

Considering these facts, we decided to add up the Latin and French loanwords, thus expressing the number of words which came in via Latin **or** French. This allows a valid comparison of the most important routes for the borrowing of Romance words into both Germanic languages considered in our study. With 86.4%, the number is 46.6% higher in Dutch than in Swedish, with 39.8%. This suggests that overall the influence of Romance languages has been considerably higher on formal Dutch than on formal Swedish.

In Figure 5 our results of the origins of loanwords are compared with the findings reported by Van der Sijs (1996) and Gellerstam (1973). We have taken Low German and High German together, since the previous investigations do

not make this distinction either. Similarly we have combined French and Latin to form a new category, which we refer to as Romance.[5] The comparison shows a striking resemblance. Apparently, the percentages of words that Dutch and Swedish borrowed from different languages are very robust and insensitive to differences between the corpora.

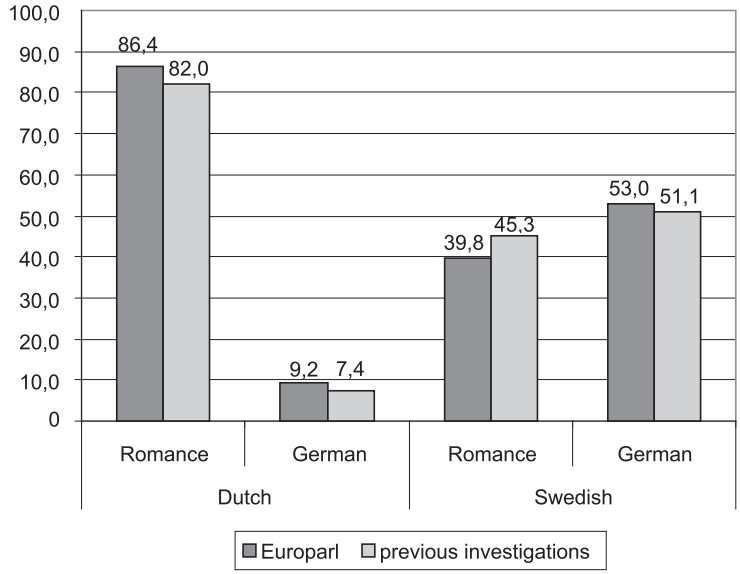

Figure 5. Origin of the loanwords in percentages in the present investigation and in earlier studies (Van der Sijs (1996) for Dutch and Gellerstam (1973) for Swedish)

Finally, in Figure 6 we consider the percentages of loanwords from specific loan-giving languages in relation to the whole lexicon, including inherited words. The figure shows clearly that the Low German loans are the only reason why the number of loanwords is higher in Swedish than in Dutch. If we left the words of Low German out of consideration, the number of loanwords would be about equal for the two languages (see dotted line in Figure 6). Considering the fact that for methodological reasons we are unable to assess the exact size of the influence of Low German on Dutch, and speculating, furthermore, that the influence of Low German on the two languages may have been rather similar, the contrast between the two languages may not be as striking as suggested above. What remains, however, is the greater influence of Romance on Dutch and the greater influence of High German on Swedish.

5. Note that Van der Sijs includes Italian and Spanish in this category and Gellerstam includes Greek.

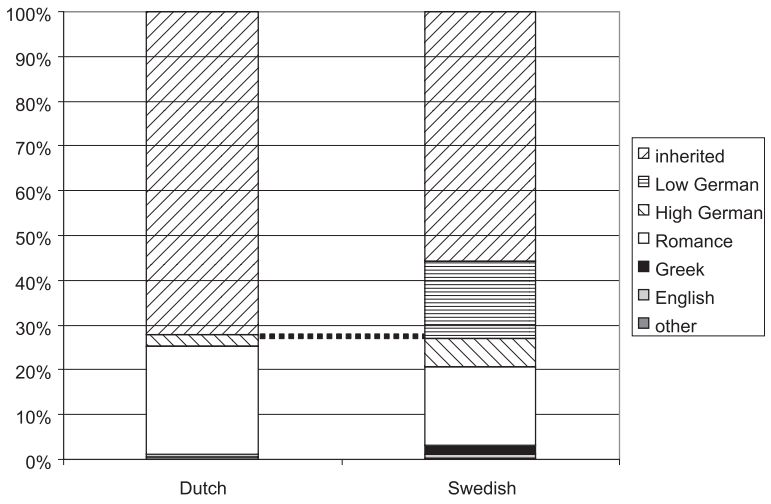

Figure 6. The percentages of loanwords from specific loan-giving languages in relation to the whole lexicon in the Europarl corpora

6. Conclusions

Our comparison of similarly constructed samples of formal spoken Swedish and Dutch revealed some clear differences in the composition of the lexicons, in spite of a common genetic background and parallels in the history of language contact. Swedish has many more loans than Dutch, namely 44.4% against 27.9%. Moreover, the nature of the loanwords differs. Within the Swedish loans there is a large compartment of Low German (38.7%), whereas most loans in Dutch have a French origin (63.8%). The differences in terms of the number and distribution of loanwords between the lexical profiles of Swedish and Dutch appear to be stable, as they were attested both in the present study and in previous studies. They can be attributed to differences in the linguistic distances between source and borrowing languages and to differences in the intensity of the contacts.

The higher number of French loans in Dutch can easily be explained. The contact between Dutch and French was more intensive and widespread than between Swedish and French. The higher number of Low German loans in Swedish needs some clarification. The intensity of the contact of Swedish and Dutch with Low German was almost identical. One would therefore expect similar numbers of Low German loans in the two languages. This is not what we found. We explained this by the fact that Low German loans are less easily identifiable

in Dutch than in Swedish due to the larger similarities between the lexicons of Middle Low German and Middle Dutch. The language contact probably resulted in frequency shifts and semantic adaptation of inherited words rather than a larger incidence of loans.

It should be noted that the reported findings pertain to formal speech. Most probably the speeches and debates in the European Parliament which formed the basis of our study were well prepared and can therefore not be characterized as spontaneous. In this sense, the nature of the *Europarl* sample may be very close to the newspaper texts analyzed in previous studies of Dutch and Swedish. As far as we know, there have been no studies of the lexical profiles of informal spontaneous speech. To fill this gap we analyzed (part of) the speech produced in the component 'face-to-face interactions' of the *Corpus Spoken Dutch*, collected between 1998 and 2004 (Van Bezooijen, Gooskens & Kürschner 2009). These are conversations about everyday topics between friends and relatives, recorded at home without an interviewer being present. The speakers are from various regions in the Netherlands and Flanders, of both sexes and of different age groups. To optimize the comparison with the findings of the present study, we applied the same selection procedure and coding. The distribution of the loanwords in this spontaneous corpus is very similar to the distribution found in the *Europarl* corpus. This confirms that the lexical profiles of Dutch are very stable and independent of the level of formality, at least as far as the most frequently used words are concerned.[6]

References

Allén, S. 1970. *Nusvensk frekvensordbok baserad på tidningstext,* 1: *Graford, Homografkomponenter.* Stockholm: Almqvist & Wiksell.

Bezooijen R. van, Gooskens, C. & Kürschner, S. 2009. Wat weet de Nederlander van de herkomst van Nederlandse woorden? *Tijdschrift voor Nederlandse Taal- en Letterkunde* 125: 324–344.

Braunmüller, K. 1995. Semikommunikation und semiotische Strategien. Bausteine zu einem Modell für die Verständigung im Norden zur Zeit der Hanse. In *Niederdeutsch und die skandinavischen Sprachen II*, K. Braunmüller (ed.), 35–70. Heidelberg: Winter.

Braunmüller, K. 2004. Niederdeutsch und Hochdeutsch im Kontakt mit den skandinavischen Sprachen. Eine Übersicht. In *Deutsch im Kontakt mit germanischen Sprachen.* H. Haider Munske (ed.), 1–30. Tübingen: Niemeyer.

6. We thank the editors and an anonymous reviewer for valuable comments on earlier versions of this paper.

Diercks, W. 1993. Zur Verwendung prä- und postmodifizierender Morpheme im Mittelniederdeutschen und in den skandinavischen Sprachen. In *Niederdeutsch und die skandinavischen Sprachen I*, K. Braunmüller & W. Diercks (eds., 161–194. Heidelberg: Winter.

Edlund, L.-E. & Hene, B. 2004. *Lånord i svenskan: om språkförändringar i tid och rum*. Stockholm: Norstedts.

Gellerstam, M. 1973. Etymologiska frekvenser i det centrala ordförrådet. *Folkmålsstudier* 23: 70–78.

Goossens, J. 2000. Herauslösung und Herausbildung des Niederländischen. In *Ausgewählte Schriften zur niederländischen und deutschen Sprach- und Literaturwissenschaft*, J. Goossens (ed.), 197–211. Münster: Waxmann.

Hellquist, E. 1980 [1922]. *Svensk etymologisk ordbok*. Tredje upplagan. Malmö: Liber. Also <http://runeberg.org/svetym/>.

Sijs, N. van der. 1996. *Leenwoordenboek. De Invloed van Andere Talen op het Nederlands*. Den Haag: Sdu & Antwerpen: Standaard.

Sijs, N. van der. 2005. *De Geschiedenis van het Nederlands in een Notendop.* Amsterdam: Bert Bakker.

Thomason, S. G. 2001. *Language Contact.* Edinburgh: EUP.

Veen, P. A. F. van der & Sijs, N. van der. 1997. *Etymologisch Woordenboek. De Herkomst van onze Woorden*. Utrecht & Antwerpen: Van Dale Lexicografie.

Wessén, E. 1960. *Våra ord. Deras uttal och ursprung. Kortfattad etymologisk ordbok.* Stockholm: Norstedt.

The impact of German on Schleife Sorbian

The use of *gor* in the Eastern Sorbian border dialect

Hélène B. Brijnen
University of Groningen

The West Slavic languages Upper Sorbian and Lower Sorbian have been surrounded by speakers of German since the early Middle Ages. The present contribution describes the sociolinguistic background and use of *gor*, a German borrowing in the dialect of Schleife, which occupies a transitional position between the two Sorbian languages. The use of *gor* in 20th century spoken Schleife Sorbian is contrasted with its use in an early 19th century source written in the same dialect, the work of the peasant-writer Hanso Nepila.

1. Introduction

The Sorbs are a West Slavic minority, inhabiting parts of the German lands of Brandenburg and Saxony. There are two main linguistic areas, each of them with its own standard language.[1] Upper Sorbian (US), which is used by a majority of Sorbian speakers, is spoken in Saxony. Lower Sorbian (LS), of which only a few thousand elderly speakers are left, is found in Brandenburg and in a few villages in Saxony.

The "smallest Slavonic nation," as the Sorbs are called by Stone (1972), has been surrounded by speakers of German since the Middle Ages. This particular situation, which is often metaphorically referred to as "a Slavic island in the German sea" (after Morfill 1883: 240; see also Stone 1972: 2), has made the Sorbs a rewarding subject for the study of German-Slavic language contact. As early as

1. The standardization of the Sorbian languages harks back to the beginning of the 18th century. To date, each language has been codified by means of normative grammars and extensive dictionaries, etc. The standard languages are used, *inter alia*, in newspapers, radio and Tvbroadcasting, literary production, school teaching and Sorbian official institutions. For further details see Stone (1972: 117 ff.) and Schuster-Šewc (1996: 249 ff.).

the nineteenth century one finds articles dedicated to the influence of German on Sorbian (e.g. Smoler 1859).

Bronisch (1862) described the German spoken in the Sorbian language area of Lower Lusatia and adjacent northern parts of Upper Lusatia, amidst a Sorbian population. He emphasizes the fact (p. 109) that, in his time, Lower Lusatia exhibited an array of different German dialects, which had become rooted in an originally Slavic environment. One may add that these dialects were also quite distinct from the Saxon dialect spoken in most of Upper Lusatia. Bronisch's study also includes a large vocabulary of Slavic loanwords in German.

However, the situation of Sorbian in the nineteenth century was very different from what it is today. The number of Sorbian speakers, estimated at 245,000 by Haupt and Smoler ([1841]1984: 9), decreased dramatically during the twentieth century. At the same time, the composition of the German-speaking population, especially that of Brandenburg, underwent radical demographic changes as a result of industrialization, mass immigration, and – after the Second World War – the formation of agricultural cooperatives. The German dialects Bronisch referred to gradually lost their distinctive character and were replaced by new intermediate forms or by standard High German (cf. Schönfeld 1990: 91–93). These facts must not be overlooked when dealing with contact phenomena between German and Sorbian. It is essential to try and specify with which variety of either German or Sorbian one is dealing, as well as the time of contact.

When comparing dialect texts and recordings from the different areas where Sorbian is spoken, it becomes clear that the influence of German is present at all linguistic levels: in the lexicon, phonology, morphology, and syntax. Some borrowings from German – for instance, the verb US *wórdować*/LS *wordowaś* 'to become' (German *werden*), which is often used in combination with a passive participle, e.g. *a wón też dej wórdować dopołdnja syty* 'and it also has to be sawn in the morning' (Spohla; SD 1963: 26) – have spread over the entire Sorbian territory. These include well known cases that have been studied by various linguists, such as the use of the demonstrative pronoun as an article (Berger 1999; in Sorbian, as in other Slavic languages, there is no article), the use of directional adverbs in combination with verbs of motion – e.g. US *won hić* 'to go out' (Toops 1992; Giger 1998; Brijnen 2000) – or the use of the neuter form of the personal pronoun *wono* as a fictive subject of subjectless sentences, e.g. US *wono so dešćuje* 'it is raining' (Faßke 1996: 74–80), etc. Borrowings from German occur in the standard languages as well, although to a lesser extent. In standard Lower Sorbian, which has been less affected by purism than standard Upper Sorbian, they are more numerous.

However, borrowings from German have only to a limited extent been studied in a cross-dialectal context (e.g. Michalk 1989; Giger 1998). Some striking

differences between dialects have been observed at the lexical level, e.g 'potato' is *běrna* (German *birne*) in standard Upper Sorbian, and *nepl* (German *Knöpfel*) in the dialect of the Upper Sorbian Catholics (cf. Schuster-Šewc 1978–1996; for an overview of the different Sorbian dialects see, for instance, Schuster-Šewc 1996: 244–248). Borrowings from German are not restricted to content words (lexical items referring to objects or actions) alone, function words may be borrowed as well (for a discussion of different types of borrowing hierarchy see Curnow 2001: 417–419; Thomason 2001: 63 ff.). The particle or adverb *gor* (German *gar* 'at all', 'very', 'really'), for instance, occurs in a majority of the dialects of the Sorbian speaking area. However, if one takes a closer look at it, one notices that, in the various dialects, *gor* does not occur with the same frequency, nor is it always used in the same way. How can this fact be explained? In what follows, I discuss the use of *gor* in what is known as the eastern border dialect and compare it to that in other dialects. First, however, I mention some facts concerning the region where the eastern border dialect is spoken, the situation of its speakers, and their contact with German.

2. Sociolinguistic background

The term 'eastern border dialect' was used by Mucke ([1891] 1965: 4–5) to designate the dialect of Schleife (Sorbian *Slěpe*), the main village of a protestant parish of the same name, situated in the eastern part of the Sorbian territory, on the border between the Upper and Lower Sorbian areas. The dialect is predominantly Lower Sorbian and emerged from contacts between speakers from these two areas, in addition to the influence of dialects spoken to the east of the Neisse. Ever since the late Middle Ages the villages where the dialect is spoken were situated amidst a scarcely populated forest on the Muscau Heath. The inhabitants were poor peasants. From the seventeenth century onwards, they belonged to the seigniory of Muscau (Sorbian *Mužakow*; modern German Bad Muskau), which was part of Saxon Upper Lusatia. After the Congress of Vienna in 1815, the area was reassigned to the Prussian province of Silesia. Under the German Democratic Republic, the Schleife region was part of the district of Cottbus. After its collapse (*die Wende*), the inhabitants opted for their incorporation into the newly formed federal state of Saxony (Freistaat Sachsen). Successive separations from either one or both of the two Sorbian core areas may have favored the maintenance of a distinct dialect, folklore, and way of life.

The dialect of Schleife is used today (2008) by fewer than one hundred elderly speakers in the seven villages belonging to the parish: Groß-Düben/*Dźěwin*, Halbendorf/*Brězowka*, Mulkwitz/*Mułkoce*, Mühlrose/*Miłoraz*, Rohne/*Rowne*,

Trebendorf/*Trjebin*, and Schleife itself. Until the beginning of the twentieth century, virtually all villagers were monolingual speakers of Sorbian. Children starting school did not understand German. Since then, the use of Sorbian in the area has declined dramatically. These developments were due, among other things, to forced Germanization during and between the two world wars, to the arrival of German expellees (*Heimatvertriebene*) from Silesia, and to the rise of the lignite industry, which led to a substantial increase of German speakers in the area. The destruction in the region caused by lignite mining is unique in Europe in its dimensions. Between 1945 and 1993, many Sorbian villages were destroyed, and their inhabitants displaced (cf. Förster 1996). Today, the process continues. The southern villages of the Schleife area are holding on to survival in the middle of a polluted industrial landscape of huge quarries, and their future existence is threatened.

Most traditional speakers of Schleife Sorbian are well over eighty years old; the few middle-aged speakers, for the most part, have undergone the influence of Upper Sorbian. The youngest generation speaks German, although there is a movement towards revitalization. From 1933 onwards, parents stopped transmitting the language to their children as a result of the repressive language policy of the Nazis; see Stone (1972: 35–36); Schuster-Šewc (1996: 257). Therefore, Schleife does not show a situation of gradual language loss from one generation to the next, but rather an abrupt transition between a generation that still speaks the language and another one that does not. Relearning the language in the region itself is possible through standard Upper Sorbian, which also traditionally has been the language of the church.

Today, Schleife has a reputation of an area where Sorbian identity is strong, as demonstrated by the practice of folkloric traditions. However, among the oldest generation, the attitude of the speakers can be ambiguous. Speakers that are willing to share their knowledge of the language and the traditions are mostly women; elderly men often refuse to talk, or are reported not to speak Sorbian, although in some cases it turned out later that both husband and wife spoke Sorbian together at home. In other cases women that know Sorbian do not speak it at home because their husbands speak only German or because they are widows. They speak with their daughters or with other women of the neighborhood, and can point out exactly who is a speaker. All confirm that in their youth in the 1920s most people in the village did not speak German, and all express their regret at the loss of language and traditions. Some women have a strong Sorbian accent when they speak German.

Today, children can learn standard Upper Sorbian in preschool and in primary school, and they can continue their studies at the Lower Sorbian Gymnasium in Cottbus/*Chośebuz*; unless they opt for the Upper Sorbian Gymnasium in

Bautzen/*Budyšin*, which is further away and difficult to reach. There is no teaching in the local dialect, which will almost certainly disappear within the next twenty years (cf. Brijnen 2004: 14–16).

3. The particle *gor*

The Sorbian particle *gor* is an element borrowed from German, where it is represented as *gar*.[2] In German, according to Drosdowski (1988: 298), the adverb or particle *gar* may have the following meanings: (1) it may be used to emphasize a negation, for instance, *gar nicht* 'not at all'; (2) in the South of the German-speaking area, it may be synonymous with *sehr* 'very', for instance, *das schmeckt gar fein* 'this is very tasty'; (3) unstressed, *gar* may be used as an emphasizing particle in rhetorical questions, for instance, *er wird doch nicht gar krank sein?* 'he wouldn't be ill, would he?'; (4) stressed, it may intensify the degree particles *zu* 'too' and *so* 'so', for instance, *ich habe es gar zu gern* 'I like it really very much'; (5) it may be used in the sense of 'maybe', for instance, *hast du es gar vergessen?* 'maybe you have forgotten it?' (see also Wahrig 1980: 1427).

3.1 The geographical distribution of *gor*

The particle *gor* has been included as a colloquial expression in Starosta's Lower Sorbian dictionary (Starosta 1999); in the Upper Sorbian standard language there is no such explicit acceptance. As a matter of fact, when one compares the various Upper Sorbian dialects, *gor* hardly appears in the dialects to the north of Bautzen (cf. SD 1967); but it does occur, although not frequently, in Radibor/*Radwor* in the Catholic area (cf. Michalk, Protze 1974), and in Rodewitz/Spree (Sorbian *Rozwodecy*) to the south of Bautzen (cf. Jentsch 1980). In the north of the Upper Sorbian area, in the border region between the two core areas, *gor* occurs more regularly. One finds it in Upper Sorbian transitional dialects with Lower Sorbian influence, for instance in Spohla/*Spale* (SD 1963), Neustadt (Michalk 1962) and in Nochten/*Wochozy* (Michalk, Protze 1967). This is only a first impression; for more precise data fieldwork in all the relevant dialects is required. Still, from the dialect samples available, one can deduce some general tendencies. In the dialects in which *gor* occurs, it appears almost exclusively in negative

2. Vowel modification is common in Sorbian borrowings of German origin. It is also possible that *gor* was taken from a German dialect featuring the change *a* > *o* (see the introduction for the fate of German dialects originally used in the Lower Sorbian region.)

sentences. Occurrences in positive sentences are rare. (See, however, the examples with *cu(jare)* discussed in Section 3.2.). The following example is from Nochten (Michalk, Protze 1967: 139):

(1) *a nět gor ma kóžd-y moped*
and now in.fact have-3SG every-NOM.SG.M moped-ACC.SG.M
'and n o w i n f a c t everyone has a moped'

Another example was recorded by Ščerba in the nowadays moribund Lower Sorbian dialect of Muskau to the east of Schleife (Ščerba [1915] 1973, Teksty, p. 11; in my transcription):

(2) *něnto chopnje-š na mnje gor sra-ć!*
now start-2SG on 1SG.ACC even defecate-INF
'now you e v e n start to defecate on me!'

The Muskau dialect is the dialect most closely related to Schleife; one of its last speakers, Emma Kral, was 94 years old in 1999.

3.2 The use of *gor* in Schleife Sorbian: *gor* in combination with *cu(jare)*

Marja Kudźelina, born in Trebendorf in 1902, was a fluent speaker of the Schleife dialect and a great storyteller. Her stories were recorded in 1963 by the late Siegfried Michalk, and are archived at the Sorbian Institute in Bautzen, which kindly put them at my disposal. In one story Kudźelina tells how once, in a field, she saw a snake (*zmija*) wearing a crown, and says:

(3) *won-a jo by-ł-a c u j a r e g o r rjan-a*
3-NOM.SG.F be-3SG be-PRF-SG.F too.very quite beautiful-NOM.SG.F
'she was e x c e p t i o n a l l y beautiful'

The combination of *gor* with *cujare* followed by an adjective to express a very high degree of something is not used by Kudźelina alone; in the data that I recorded in Trebendorf in 1999, one of the youngest speakers and a distant relative of Kudźelina says:

(4) *mój brat móžo g o r c u j a r e*
my-NOM.SG brother-NOM can-3SG quite too.very
dobr-y tej wari-ć
good-ACC.SG.M tea-ACC cook-INF
'my brother can make e x t r e m e l y good tea'

The form *cujare*, composed by *cu* and *jare*, is interesting in itself.[3] *Cu* is a borrowing (German *zu*, see above), which, in the meaning 'too', 'excessively' (as well as in the meaning 'towards') is quite common throughout the Sorbian territory.[4] *Jare* is the normal expression for 'very' in the eastern border dialect; in Upper Sorbian, it occurs as *jara*.

The lexicalized combination *cujare* is a particularity of the Eastern Border dialect. It is also found, as *cujara*, in the neighboring Upper Sorbian village of Neustadt/*Nowe Město*.[5] Both forms function as emphatic alternatives of *jare/jara*; for instance, in *cujare dobry twarog* 'delicious cheese' (Kudzélina), and in *cujara wělka hica* 'excessive heat' (Michalk 1962: 365).

In the dialect of Schleife, *cujare* can be accompanied by *gor* for further emphasis. In *gor cujare dobry tej* (see above), *gor* is stressed and precedes *cujare*, which is comparable to German *gar zu* (cf. Drosdowski 1988: 298; Klosa 2001: 597). In *cujare gor rjana*, *gor* is unstressed and occurs in a position where it would not occur in German. This unusual use of *gor* in combination with *cujare* made it worthwhile to investigate the other ways in which *gor* is used in Schleife and elsewhere.

3.3 *Gor* in the writings of Hanso Nepila (1761–1856)

In the villages where the dialect of Schleife is spoken today, apart from the examples with *cujare* given above, one finds *gor* in negative sentences. When preceding the verb form, *gor* emphasizes the negation; for example *gor nic* 'nothing at all', *gor njewě* 'doesn't know at all'; when following the verb form and unstressed, *gor* may function more like a discourse marker; for example (Lena Šprjejcowa, Rohne 1999) *to njewěm gor* 'that I don't know actually'. In all cases, *gor* occurs in a negative context. Neither in Schroeder 1958, nor in Michalks recordings from the 1960s, nor in my own fieldwork from 1999, did I find positive sentences with *gor* like the ones from Nochten and Muskau above (cf. 3.1).

This is all the more surprising because in the original writings of the peasant Hanso Nepila from Rohne, which date from the first decades of the nineteenth century (cf. Brijnen 2004), the particle *gor* is highly frequent, both in negative and

---

3. This form is sometimes found without the vowel *u*; for example, in *cjare fajn* 'very smart' (Schroeder 1958: 87).

4. For example, in LS *cu rědnje* 'too beautiful' (SD 1965: 26), and in US *cu wjele* 'too much' (Jentsch 1980: 223).

5. Neustadt is situated at a distance of only a few kilometers to the southwest of Mulkwitz and Mühlrose.

in positive sentences. Nepila kept a diary in his native Schleife dialect, in which he wrote about domestic affairs and quarrels. Although Nepila's writings contain some archaisms, such as verbal synthetic past forms, which are lost today, present-day speakers easily identify his language as their own.

In negative sentences Nepila uses *gor* in much the same way as is it is done today; for example (the examples given in my transcription):

(5) *a na to won-a nje-praja-š-o g o r n i c*
and to DEM.ACC.SG.N 3-NOM.SG.F NEG-say-IPF-3SG at.all nothing
'and to that she said n o t h i n g a t a l l'

In positive sentences *gor* occurs with *cujare* just as it does today; e.g.

(6) *won-a jo až baldy g o r c u j a r e wjelik-a*
3-NOM.SG.F be.3SG until soon quite too very big-NOM.SG.F
'it is almost t o o big'

but it also appears in other positive sentences, for instance, in combination with the particle *tak* 'so' (cf. German *gar so*), with or without *jare*; e.g.

(7) *ty južon ma-š g o r t a k w j e l e pjenjez*
2SG already have-2SG quite so much money.GEN.PL
'you already have s u c h a l o t of money'

Furthermore, *gor* may precede or follow adverbs or adjectives to express 'very, really, quite'; e.g.

(8) *a mi jo se też to g o r dź i w n j e zda-ł-o*
and 1SG.DAT be.3SG REFL also it quite odd.ADV seem-PRF-SG.N
'and it also seemed q u i t e o d d to me'

(9) *a też někoter-e gałuz-e s t a r e g o r*
and also some-NOM.PL branch-NOM.PL old-NOM.PL very
'and also some r a t h e r o l d branches'

Gor may also emphasize a noun, as in (10):

(10) *ten jo g o r p j e n j e z - e mě-ł*
DEM.NOM.SG.M be.3S really money-ACC.PL have-PRF.SG.M
'he s u r e had m o n e y'

(11) *a g o r ć m a bě-š-o*
and really darkness-NOM.SG be-IPF-3SG
'and it was r e a l l y dark'

or it may combine with a verb, e.g.:

(12) *wej derja-ł-ej se gor sroma-ć*
2DU ought.to-PRF-DU REFL really be.ashamed-INF
'you both ought to be really ashamed'

Gor regularly occurs in ironic expressions; e.g.

(13) *jo jo tedym by gor kjarla by-l-i*
yeah yeah then COND quite man.NOM.SG be-PRF-PL
'yeah, yeah, then you would be quite a man'

Furthermore, *gor* frequently appears in expressions of time or size, often in combination with the adjective *chytry* 'considerable, substantial, quite a'; for example, *gor dobre běrtyl štundy* 'a good quarter of an hour'; *gor chytro wjele (togo drjowa)* 'quite an amount (of wood)'; *gor tajka chytra łuža* 'such an enormous puddle'. The sequence *gor chytru chwilu* 'quite a while' occurs so often that it can be considered a fixed expression in Nepila.

Another such fixed combination is *jeno* 'only' + *gor*; for example, *jeno gor to nejgorše* 'only the worst'; *jeno gor te ćeńke* 'only the thin ones'. *Jeno gor* may be further emphasized by *rowno* 'just, quite, exactly':

(14) *jeno gor rowno jěšći ten kužd-y*
only quite just yet DEM.ACC.SG.M every-ACC.SG.M
dźeń nic
day.SG.ACC nothing
'only not quite everyday yet'

Finally, Nepila uses *gor* as an equivalent of German *sogar* ('even'), as recorded by Ščerba in Muskau (see above); e.g.

(15) *jen-u aby tež gor dwě*
one-ACC.F or also even two.ACC.F
'one or even two'

(16) *a wóter-e raz-y gor tśi*
and some-ACC.PL time-ACC.PL even three-ACC
raz-y za tydźeń
time-ACC.PL per week.ACC.SG
'and sometimes even three times a week'

and as an equivalent of German *ganz* ('completely'), e.g.

(17) *won-a mi potom gor hynak spiwa-š-o*
3-NOM.SG.F 1SG.DAT then completely different sing-IPF-3SG
'then she would play a completely different tune to me'

4. Discussion

Why would *gor* be more frequent today in Upper Sorbian dialects with a Lower Sorbian influence, than in dialects where such an influence is not found? How can one explain the striking difference between the use of *gor* by Nepila in the nineteenth century and its use by the present-day speakers of the Schleife dialect?

Insofar as the first question is concerned, one might be inclined to think that the character of the language communities at issue and a difference in the amount of ethnic awareness could play a role. Due, among other things, to a more frequent use of Sorbian in church, feelings of ethnic identity in the Upper Sorbian core area may be much stronger than in both the Lower Sorbian area and the border region, where most speakers have an ambivalent attitude towards their native language. One might also consider the fact that in the border area the last few speakers are more exposed to German than in Upper Sorbian villages where the percentage of Sorbian speakers is higher, and the influence of the standard language stronger. Although these facts are relevant for the preservation of the language, I do not believe they answer the question relating to the present-day distribution of *gor*. In Upper Sorbian dialects, even those of the Catholic core area where entire villages are Sorbian-speaking, German borrowings are equally abundant. They include particles and discourse markers, such as *also* 'well'; *blows* 'only'; *doch* 'all the same'; *ebm* 'just, indeed'; or *jemol* 'once' (German *also*, *bloß*, *doch*, *eben*, *einmal*, respectively). There seems to be no special hierarchy that favors the borrowing of words referring to concrete objects or actions, rather than function words. Instead, the distribution of *gor* suggests a preference for the borrowing of words matching the phonological structure of the language at issue. The hypothesis that "borrowing is more easily achieved if there is something similar between the two languages" (Curnow 2001: 424) appears to be confirmed by *gor*. In this particular case, the similarity consists in the presence of a consonant [g] in the phonological system of the dialects where *gor* is frequent.[6]

As for the second question formulated above, the less restricted use of *gor* in Nepila, in comparison to present-day speakers of the Schleife dialect, brings up the issue of the amount of exposure to German that Nepila experienced in his time. Did Nepila actually speak German himself? These questions, unfortunately, cannot be answered with certainty today. However, if one considers that even at the beginning of the twentieth century speakers of Schleife Sorbian needed an interpreter when dealing with the authorities or when selling their products on the market in Spremberg/*Grodk*, one may safely assume that for Nepila the

6. Lower Sorbian, including the Schleife dialect, has [g]; in Upper Sorbian *g > h (see also Stone 1972: 92ff.).

situation was not much different. However, what sort of German was spoken in Nepila's neighborhood? It is remarkable that Nepila's speech lacked several borrowed function words that are frequent in the dialect as it is spoken today. His writings lack, for instance, the form *ganc* 'very, completely' (German *ganz*), as in present-day Schleife Sorbian *ja som była ganc sama wóstała* 'I stayed behind completely alone' (Marta Mrosk, Trebendorf 1999; compare Nepila *gor hynak spiwašo*, example (17)). This leads one to think that the explanation must be sought in the variety of German that surrounded Nepila in his time. Many loanwords in Nepila – for instance, *pług* 'plough'; *punt* 'pound'; *sejpa* 'soap' (High German *Pflug*, *Pfund*, *Seife*) – clearly show the influence of Low German. In Brandenburg, this influence has decreased in the course of the nineteenth and twentieth centuries (cf. Schönfeld 1990: 91–93). Nepila's vocabulary can help to assess the former southward extension of Low German. Furthermore, it would be interesting to trace the use of particles in Central and East Low German at the beginning of the nineteenth century. However, the combination of *gor* with *cujare*, as in (3) *cujare gor rjana*, seems to be the result of an internal development in Schleife Sorbian.

By comparing the Sorbian dialects and the Germanisms they contain, synchronically as well as diachronically, one can learn as much about the German language and its history as about Sorbian. This challenging field deserves to be explored further but extends beyond the scope of this article. Similarly, the study of Lower Sorbian syntax, which has received very little attention so far, deserves to be developed as well.

References

Berger, T. 1999. Die Gebrauchsbedingungen des bestimmten Artikels im älteren Obersorbischen. *Lětopis* 46: 7–23.

Brijnen, H. B. 2000. German influence on Sorbian aspect: The function of directional adverbs. In *Languages in Contact* [Studies in Slavic and General Linguistics 28], D. G. Gilbers, J. Nerbonne & J. Schaeken (eds), 67–71. Amsterdam: Rodopi.

Brijnen, H. B. 2004. *Die Sprache des Hanso Nepila. Der niedersorbische Dialekt von Schleife in einer Handschrift aus der 1. Hälfte des 19. Jahrhunderts* [Schriften des Sorbischen Instituts 35]. Bautzen: Domowina-Verlag.

Bronisch, K. W. 1862. Grundzüge der deutschen Mundart, welche inmitten der sorbischen Bevölkerung und Sprache der Niederlausitz und in den nördlichen Teilen der Oberlausitz gesprochen wird. *Neues Lausitzisches Magazin*: 108–195.

Curnow, T. J. 2001. What language features can be borrowed? In *Areal Diffusion and Genetic Inheritance. Problems in Comparative Linguistics*, A. Y. Aikhenvald & R. M. W. Dixon (eds.), 412–436. Oxford: OUP.

Drosdowski, G. et al. 1988. *Duden, Stilwörterbuch der deutschen Sprache*. Mannheim: Dudenverlag.

Faßke, H. 1996. *Sorbischer Sprachatlas,* Vol. 15. *Syntax.* Bautzen: Domowina-Verlag.

Förster, F. 1996. *Verschwundene Dörfer. Die Ortsabbrüche des Lausitzer Braunkohlenreviers bis 1993* [Schriften des Sorbischen Instituts 8]. Bautzen: Domowina-Verlag.

Giger, M. 1998. Zu Lehnübersetzungen und Entlehnungen deutscher postponierbarer Präverbien in sorbischen Dialekten. In *Schweizerische Beiträge zum XII. Internationalen Slavistenkongress in Krakau, August* 1998 [Slavica Helvetica 60], J. P. Locher (ed.), 129–170. Frankfurt: Peter Lang.

Haupt, L. & Smoler, J. E. [1841]1984. *Volkslieder der Sorben in der Ober- und Niederlausitz*. Bautzen: Domowina-Verlag.

Jentsch, H. 1980. *Die sorbische Mundart von Rodewitz/Spree*. Bautzen: Domowina-Verlag.

Klosa, A. et al. (ed.) 2001. *Duden. Deutsches Universal-Wörterbuch.* Mannheim: Duden.

Michalk, S. (= Michałk, F.). 1962. *Der obersorbische Dialekt von Neustadt*. Bautzen: Domowina-Verlag.

Michalk, S. (= Michałk, F.). 1989. Zur Frage der Häufigkeit deutscher Lehnwörter in den sorbischen Dialekten. In *Zbornik razprav iz slovanskega jezikoslovja*, F. Jakopin (ed.), 163–174. Ljubljana: Slovenska Akademija Znanosti in Umetnosti.

Michalk, S. & H. Protze 1967. *Studien zur sprachlichen Interferenz* I: *Deutsch-sorbische Dialekttexte aus Nochten, Kreis Weißwasser*. Bautzen: Domowina-Verlag.

Michalk, S. & Protze, H. 1974. *Studien zur sprachlichen Interferenz* II: *Deutsch-sorbische Dialekttexte aus Radibor, Kreis Bautzen.* Bautzen: Domowina-Verlag.

Morfill, W. R. 1883. *Slavonic Literature*. London.

Mucke, E (= Muka, A.). [1891] 1965. *Historische und vergleichende Laut- und Formenlehre der niedersorbischen (niederlausitzisch-wendischen) Sprache*. Leipzig: Zentral-Antiquariat der DDR.

Schönfeld, H. 1990. East Low German. In *The Dialects of Modern German. A Linguistic Survey,* C. V. J. Russ (ed.), 91–135. London: Routledge.

Schroeder, A. 1958. *Die Laute des wendischen (sorbischen) Dialekts von Schleife in der Oberlausitz. Lautbeschreibung*. Tübingen: Niemeyer.

Schuster-Šewc, H. 1978–1996. *Historisch-etymologisches Wörterbuch der ober- und niedersorbischen Sprache.* Bautzen: Domowina-Verlag.

Schuster-Šewc, H. 1996. *Grammar of the Upper Sorbian Language. Phonology and Morphology* [Lincom Studies in Slavic Linguistics 3], trans. by G. H. Toops. Munich: Lincom.

SD = *Sorbische Dialekttexte* 1963. Faßke, H. & S. Michalk (eds): *I. Spohla, Kreis Hoyerswerda*. Bautzen: Domowina-Verlag.

SD= *Sorbische Dialekttexte.* 1965. Faßke, H. & H. Jentsch (eds): *III Schmogrow, Kreis Cottbus*. Bautzen: Domowina-Verlag.

SD = *Sorbische Dialekttexte* 1967. Faßke, H. & S. Michalk (eds): *V. Klix, Kreis Bautzen mit Spreewiese, Salga und Göbeln*. Bautzen: Domowina-Verlag.

Smoler, J. E. 1859. Přichodny čas serbskeho słowjesa. *Časopis Maćicy Serbskeje* XII (1): 7–14.

Starosta, M. 1999. *Dolnoserbsko-nimski słownik. Niedersorbisch-deutsches Wörterbuch.* Bautzen: Domowina-Verlag.

Stone, G. 1972. *The Smallest Slavonic Nation. The Sorbs of Lusatia.* London: The Athlone Press of the University of London.

Ščerba, L. V. [1915] 1973. *Vostočnolužickoe narečie (Der ostniedersorbische Dialekt)*. Bautzen: Domowina-Verlag.

Thomason, S. G. 2001. *Language Contact.* Edinburgh: EUP.

Toops, G. H. 1992. Lexicalization of Upper Sorbian preverbs: Temporal-aspectual ramifications and the delimitation of German influence. *Germano-Slavica* VII (2): 3–22.

Wahrig, G. 1980. *Deutsches Wörterbuch.* München: Mosaik Verlag.

Detecting contact effects in pronunciation

Wilbert Heeringa, John Nerbonne and Petya Osenova
Variationist Linguistics, Meertens Institute, Amsterdam /
University of Groningen / IPOI, Sofia

We investigate language contact effects between Bulgarian dialects on the one hand, and the languages of the countries bordering Bulgaria on the other. The Bulgarian data comes from Stojkov's Bulgarian Dialect Atlases. We investigate three techniques to detect contact effects in pronunciation, the phone frequency method and the feature frequency method, both of which are insensitive to the order of phonological segments within words, and also Levenshtein distance, a word-based method which is order-sensitive. We also examine pronunciation effects under the hypothesis that pronunciation influences should be strongest as one approaches the border of a country which speaks the putatively influential language. The study aims to contribute to the development of more exact tools for studying language contact.

1. Introduction[1]

Although computational techniques have recently enabled large scale investigations of language varieties (Nerbonne & Kretzschmar 2006, and references there), little computational attention has to-date been paid to techniques for assaying language contact effects. Heeringa, Nerbonne, Niebaum, Nieuweboer & Kleiweg (2000) studied Dutch-German contact in and around the German county

1. The authors would like to thank Kiril Simov for help in digitizing the data; Luchia Antonova for comments on the IPA conversion, for the selection of the Bulgarian sites and for general recommendations on Bulgarian dialectology; Christine Siedle for her help with geographical coordinates and the maps; and Peter Kleiweg for software and his quick reactions on software questions. We also owe thanks to the audience at *Language Contact in Times of Globalization* for very useful discussion and suggestions on a preliminary version of this paper. We especially thank Peter Houtzagers and Muriel Norde for their valuable remarks. This work is funded by NWO, Project Number 048.021.2003.009, P. I. J. Nerbonne, Groningen, and also a grant from the Volkswagenstiftung "Measuring Linguistic Unity and Diversity in Europe", P. I. E. Hinrichs, Tübingen.

Bentheim. They found that dialects at the Dutch side of the border have become more Dutch while the German dialects have become more German. Measurements were made with the use of Levenshtein distance, which measures pronunciation differences between pairs of words, preferably pairs of cognates.

For centuries Bulgaria has been in intensive contact with its neighboring countries. This contact includes relations not only in the areas of politics and economics, but also among languages. In this paper we compare dialects throughout Bulgaria to the five standard languages on its borders, viz., Macedonian, Serbian, Romanian, Greek and Turkish. We use a design intended to capture areal effects in language contact (Kurath 1972). We hypothesize that the areal spread of linguistic features should result in gradients of increasing similarity between the various dialects and each of the putative sources of contact effects. For example, in the case of Romanian, this predicts that varieties closest to the Romanian border will be most similar to Romanian, and those furthest away most dissimilar. The thesis that pronunciation should be subject to mixing effects stands in contrast to the general position of Balkanologists, who regard pronunciation as little affected by widespread contact (Birnbaum 1965). Our study uses a simple model of geography (effectively, just linear distance) and studies whether phonological similarity is related to it.

Naturally we need to operationalize the notion 'phonologically similar' in order to do this. We cannot rely exclusively on the human observations to adjudge phonological similarity since we need a method that can be applied to large amounts of material automatically, i.e. a computational technique. The Bentheim study used Levenshtein distance, a technique which aligns corresponding segments of the words to be compared, and sums the differences between the segments. But Levenshtein distance is sensitive to the order of segments in words, and insensitive to differences in segments that do not correspond. If we consider the example of the spread of uvular /r/ in the languages of Europe (Chambers & Trudgill 1998 [1980], §11.4), it is clear that we notice changes even when they do not involve corresponding words. The uvular /r/ is present e.g. in the German word [ʀaux, ʁaux] 'smoke', even though it is completely absent in the nearest French equivalent *fume* /fym/ (and even though French is the source of the uvular /r/, as scholars agree). We are therefore cautious about applying Levenshtein distance to materials from very different languages.

We therefore also consider two other corpus-based techniques where the difference between two dialects is equal to the sum of phones or, alternatively, features frequency differences of the respective corpora. The phone frequency method (PFM) was introduced by Hoppenbrouwers & Hoppenbrouwers (2001) and the feature frequency method (FFM) was firstly introduced by

Hoppenbrouwers & Hoppenbrouwers (1988), but described in its most mature form in Hoppenbrouwers & Hoppenbrouwers (2001). In a nutshell, PFM compares two languages or language varieties by counting how many tokens there are of each phoneme in comparable corpora. FFM is a step more abstract, counting how many tokens there are of segments with specific values for given phonological features. Both of them seem poised to detect interlanguage effects that might escape Levenshtein distance.

The structure of the paper is as follows: the next section provides some background on Bulgarian dialectology. Section 3 focuses on the data source and the preparation of the data. In Section 4 the dialect distance metrics are explained and the procedure to measure the geographical course of the influence of surrounding languages to the Bulgarian dialect continuum. Section 5 discusses the results, and postulates that one enigmatic aspect of the present analysis has its roots in earlier patterns of settlement in Bulgaria. Section 6 sketches conclusions and prospects for further work along these lines.

2. Background: Bulgarian dialectology

Since we shall test a hypothesis about language contact by examining whether Bulgarian dialects become more and more similar to contact languages as one approaches the borders, we review the basic facts of Bulgarian dialectology here, focusing on pronunciation. It will be important later to conclude that the measurements we are making do not contradict what is known about Bulgarian dialects. Our presentation of this background follows Stojkov (2002).

There is a major east/west division following the pronunciation of the old Bulgarian vowel 'yat' (in Bulgarian: 'ят'). In western Bulgarian dialects 'yat' has only the reflection /e/, e.g. *bel* 'white' – *beli* 'white-pl', while 'yat' in eastern dialects shows both reflections, /e/ and /ja/, e.g. *bʲal* 'white' – *beli* 'white-pl'. This single characteristic does not by itself distinguish the dialects consistently, but it remains quite important.

The various historical developments of the old Bulgarian 'big nosovka' (in Bulgarian: '*голяма носовка*'), a nasal vowel, divide Bulgarian dialects into five groups: ə-dialects (Northeastern and Northwestern Bulgaria and the eastern part of Southeastern Bulgaria); a-dialects (Western Bulgaria and the eastern dialect of Pirdop); ʊ-dialects (the Rodopi mountain); æ-dialects (the Teteven region and two villages in Eastern Bulgaria, Kazichino and Golitsa); and u-dialects (Western Bulgarian areas near the Bulgarian-Serbian border). This classification is admirably simple but also encounters numerous exceptions.

Morphological and lexical research shows Bulgaria to be divided into a central part (Northeastern and Central Bulgaria) and a peripheral part (Northwestern, Southwestern and Southeastern Bulgaria; Stojkov, 2002, p. 93).

Because of the instability and conflicting nature of various linguistic criteria Stojkov (2002) suggests a classification of Bulgarian dialects which respects geographical continuity, as well. In his standard work he distinguishes six, rather than five areas, concluding first that Bulgarian dialects are not separated categorically, but rather form a continuum. Second, there is a central (typical) area as well as peripheral (transition) areas among Bulgarian dialects. Third, Stojkov agrees with traditional scholarship that the most striking distinction of Bulgarian dialects is between East and West along the 'yat' border. In Figure 1 the six most significant geographical groups of Bulgarian dialects are shown as presented in Stojkov (2002, p. 416).

The vertical lines represent Moesian dialects; the horizontal lines represent Balkan dialects, the broken slanting lines – Southwestern dialects, the crosses – Northwestern dialects. The thick broken line represents the 'yat' borderline that divides the dialects into two major groups: Western and Eastern. The nearly horizontal slanting lines on the left side show transitional zones, and the steeply slanting lines at the bottom of the map represent the Rupskian (Rodopian) dialects.

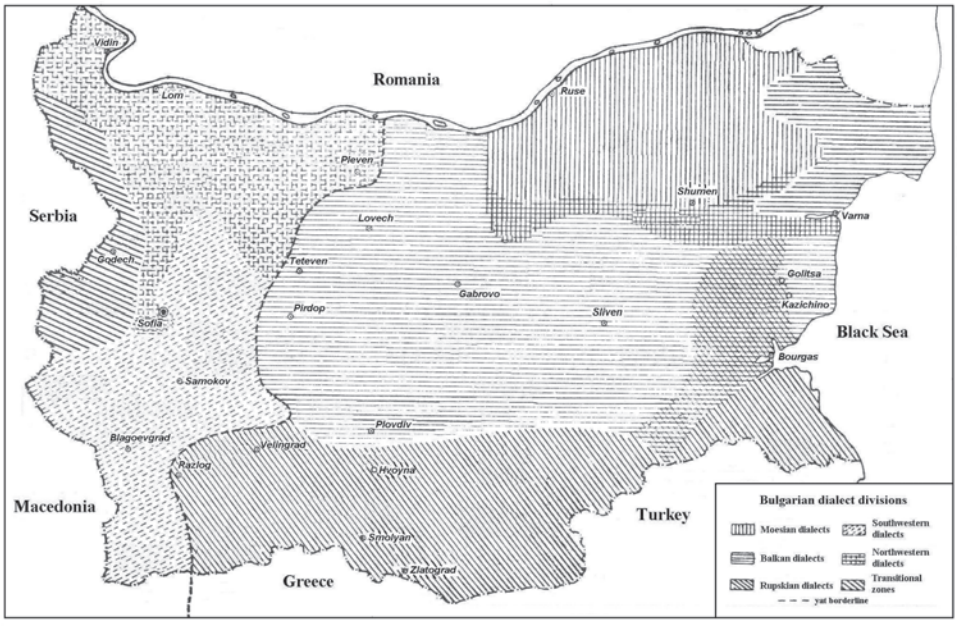

Figure 1. The map of Bulgarian dialect divisions as presented in Stojkov (2002, p. 416)

For the purposes of this paper it is important to note that there are dialect divisions which indeed correspond to the various "peripheral areas". Naturally, this does not mean that these areas are therefore more similar to the languages spoken on the other side of the border, this is something we shall test. As we shall show below, the areas of similarity are in any case more diffuse than the division into areas suggests (see Section 4, Figures 3, 4 and 5).

3. The data

We consider in turn the sources of our data and the selection we made, its preparation, and its conversion to digital form.

3.1 Sources

The data was digitized from the four volumes of Bulgarian dialect atlases which cover the entire country. These volumes are described in Stojkov (2002) and also Osenova, Heeringa and Nerbonne (forthcoming), and we shall repeat only the most important information for our purposes here. The atlases were compiled over a period of thirty years by various fieldworkers, who transcribed consistently into a broad phonetic transcription. Fieldworkers did not rely on single informants, but instead used several, and attempted to elicit material indirectly in extensive interviews, rather than via direct questions.

We extracted words from these atlases which we then compared in pronunciation. Our method (described below) relies on transcriptions of entire words, which we took from the atlases as best we could. Where we needed (infrequently) to extrapolate, we always did this conservatively, i.e. using no additional phonetic detail.

The sites sampled in the atlases were all exclusively ethnic Bulgarian populations regardless of geography. We speculate that the atlas designers chose only such sites because they were interested in the historical roots of Bulgarian. Whatever the reason, the selection is clearly suboptimal for the purpose of gauging contact effects; indeed, it seems better designed to hide contact effects rather than document them. However, instead of giving up in recognition of this problem, we choose to forge ahead, reasoning that long-standing effects of the sort we are interested in should not occur only in ethnically heterogeneous settlements. Further, we suspect the effects of restricting attention to ethnically homogeneous towns and villages should not confound the study, since it affects all areas in roughly the same way. But it remains the case that the sites sampled in the atlas certainly under-represent the degree of contact influence in the country.

3.2 Sites

In Stojkov's Bulgarian Dialect Atlases data from 1682 sites is available. We use a subset of 488 sites which were selected with respect to two main criteria: maximally complete coverage of the area covered by the atlas, and a representative number of varieties and sub-varieties. We would have preferred using sites selected randomly from a regular grid throughout Bulgaria, but there were no collection sites in large stretches of the country, which explains the patchy impression of the map. The distribution of the 488 sites is shown in Figure 2.

When studying the influence of a particular language on the Bulgarian dialects, and especially the course of the influence in the Bulgarian dialect landscape, we need to measure the shortest geographic distances to the border of the country in which that language is spoken. We measured these distances manually using a paper

Figure 2. The distribution of the 488 Bulgarian sites selected. We use a subset of 50 dialects to study the relationship between geography and language contact with languages of bordering countries. These sites are represented by circles. The boundary lines indicate administrative divisions, not dialect areas.[2]

2. Here the older administrative division is presented (valid up to 1997). We prefer this representation, because the areas are few, and thus easily detectable. The new division includes 28 regions. The interested reader is referred to: http://bg.wikipedia.org/wiki/Административно_деление_на_България.

map and a ruler. Because this turned out to be time-consuming, we restricted the analysis to a subset of 50 representative varieties (from the original 488), which were scattered as regularly as possible. The 50 sites are represented by circles in Figure 2. For clarity, we should note that in this paper we use the 488 sites for calculating and visualizing dialect distances compared to the standard languages (see Sections 5.1 and 5.2 and Figures 3, 4 and 5), and we use the selection of 50 sites for the regression analysis (see Section 5.3) aimed at detecting contact effects.

3.3 Words and conversion

We digitized a set of 54 words, which turned out not to be instantiated at every site, but which includes a subset of 36 words that were instantiated in all the atlas volumes. This differentiation of two sets arose because, as noted above, the lexical material differs across the four atlases.

The digitization step involved transliterating from a Bulgarian system of phonetic transcription into IPA, which was processed in its computerized form, X-SAMPA. We include two tables in an appendix to show how we interpreted the Bulgarian phonetic transcription system in terms of equivalents in the *International Phonetic Alphabet* (IPA 2003).

Table 1 in the Appendix provides the list of 36 words that were common to all of the 488 sites selected from the atlases. The phonetic transcriptions of the standard Bulgarian, Macedonian, Serbian, Romanian, Greek and Turkish pronunciation are given. The transcriptions are the same as used for the experiments in this paper. Osenova, Heeringa and Nerbonne (forthcoming) discuss the word sample and its properties in more detail.

The words in Table 1 represent many of the most important phonetic features of Bulgarian varieties. They reflect the following phenomena:

1. the reflections of 'yat' in different phonetic contexts (stressed and unstressed, word-finally, after fricatives, etc.): [bʲala, ˈbɛli, ˈgrɛʃka, mlɛˈkar, ˈvətrɛ, vɛnˈʧilo]
2. the reflections of the etymological 'ja': [ˈjazdi] or [ˈjɛzdi], [poˈlʲana] or [poˈlʲɛna], [guˈlʲaj] or [guˈlɛj], [dɛn] or [dɛnʲ]
3. nonpalatal-semipalatal-palatal distinction word-finally: [sol] or [solʲ] or [solˈ], [pət] or [pətʲ] or [pətˈ], [kon] or [konʲ] or [konˈ], [dɛn] or [dɛnʲ] or [dɛnˈ][3]

3. The IPA system (revised to 2005) provides a diacritic for palatalized segments, but does not distinguish between semipalatalized and palatized segments. In the Bulgarian atlas, however, this distinction is made. Here we add a superscript j to semipalatalized segments (e.g. [tʲ]) and a ˈ to palatalized segments (e.g. [tˈ]). When processing the data, we do not yet process the semipalatalized diacritic, but ignore it.

4. the realizations of 'schwa' under stress: ['bəʧva] or ['boʧva] or ['baʧva] etc. The same for other words: ['zəlva, sən, 'tənko, oti'ʃəl, do'ʃəl]
5. the realizations of the nasal vowel: [zəb] or [zob] or [zab] etc. Similarly for other words: ['kəʃta, 'səbota]
6. the metatheses 'əl-lə' and 'ər-rə': [ʒəlt] or [ʒlət], [gərb] or [grəb]
7. the realizations of various vowels in different contexts: ['ovʧɛ] or ['ovʧo], [klʲuʧ] or [kliʧ]
8. the reduction of the open vowels in unstressed position: [mlɛ'kar] or [mli'kar], [vɛn'ʧilo] or [vin'ʧilo], etc.

3.4 Contact language material

To compare the pronunciations in the contact languages, we used the most frequent lexicalization of the concepts used in the word list above. We sought these for each of the four contact languages examined: Macedonian, Serbian, Romanian, Greek and Turkish. Macedonian and Serbian are closely related South Slavic languages, while Romanian belongs to the Romance language family, Greek to Greek and Turkish to Turkic. This appears to be unfortunate from the point of view of language contact studies, as it will be impossible to separate genealogical influence (stemming from the common historical source of the Slavic languages) from contact influence in the case of Macedonian and Serbian, but it is quite fortunate in that we can use the well-known proximity of Bulgarian to Macedonian and Serbian as a test of how well the different candidate techniques are working.

The set of 36 words which we used for the comparison comprises almost exclusively words of Slavic origin. Only two loanwords are present: 'pocket' and 'pot', both from Turkish. The nearest equivalents in Macedonian and Serbian were obtained from Bulgarian experts on these languages on the basis of the Bulgarian words. The nearest equivalents in Romanian, Greek and Turkish were obtained by asking native speakers of these languages for the nearest equivalent, using English translations as a basis for comparison.

It was naturally difficult at times to settle on a single closest word for a given concept. For example, Turkish has two words for the concept 'mistake'; Romanian two words for 'feast'; and Serbian two words for 'cup, glass' (as does English). In all these cases, both words were used and the differences averaged. In cases of morphosyntactic asymmetry, in which single lexical items in Bulgarian were closest to multiword lexical items (e.g. 'ride' in Turkish) we encode the sequence of words and used that as a basis of comparison.

4. Methods

4.1 Measuring linguistic distances

4.1.1 *Phone frequency method*

Hoppenbrouwers and Hoppenbrouwers (2001, p. 1) describe an experiment in which languages were compared on the basis of phonetic texts from *The Principles of the International Phonetic Association* (IPA 1949). In this IPA pamphlet the fable 'The North Wind and the Sun' is produced in 51 languages and rendered in phonetic transcription. For each text the frequencies of phones are determined. Since not all samples have the same size, relative frequencies are used. The distance between two languages is equal to the sum of the absolute values of the differences between the corresponding (relative) phone frequencies.

Even though the difference between palatalized and non-palatalized consonants is represented only through the use of the diacritics (see examples, Appendix Table 1), palatal and nonpalatal consonants were regarded as distinct when we counted the phones. Applying this method to our material (488 dialects, standard Bulgarian, Macedonian, Serbian, Romanian and Turkish) 69 different phones were found.

4.1.2 *Feature frequency method*

The phone frequency method introduced above does not take into account that for example the [i] and [ɪ] are more similar to each other than the [i] and [ɑ]. Therefore Hoppenbrouwers and Hoppenbrouwers (1988) developed the feature frequency method. Using this method, each phone is described by a range of binary features. For example the feature *round* is set to 1 when a vowel is rounded (e.g. the [y]) and set to 0 if a vowel is not rounded (e.g. [i]). The feature *voiced* is set to 1 when a consonant is voiced (e.g. [v]) and set to 0 if the consonant is not voiced (e.g. [f]). If we have a corpus of 36 phonetic transcriptions per variety, for each feature we count the number of segments for which that feature is marked positively. We count the number of rounded sounds, the number of voiced sounds, etc. The frequencies are divided by the total number of phones in the corpus to obtain relative frequencies. We calculate the distance between two languages as the sum of the differences between the corresponding feature frequencies.[4]

When defining phonetic segments in terms of features, one has to choose the right features. Hoppenbrouwers and Hoppenbrouwers (2001) used a modified

4. Hoppenbrouwers and Hoppenbrouwers (1988) give several alternatives for calculating the difference of two feature frequency histograms. It is beyond the scope of this article to discuss them fully.

version of *The Sound Pattern of English* (SPE) (Chomsky & Halle 1968). We used the system of Almeida & Braun (1986), since this system is directly based on the well-known IPA system. When using this system, we separate vowels and consonants. It means that vowel feature counts are divided by the number of vowels in the corpus, and consonant feature counts are divided by the number of consonants in the corpus. The vowel features are listed in Table 2 and the consonant features are listed in Table 3.

We converted the IPA-inspired Almeida & Braun system to a binary system (like SPE). The binary system chosen is designed to avoid the obscuring effects of multivalued systems in which contrasting differences may be neutralized. We illustrate the danger with a small example. Assume that one variety has one front vowel and one back vowel. The mean value will be equal to (1+3)/2 = 2. Another variety with two central vowels would have a value of (2+2)/2 = 2. In this way it looks if the two varieties do not differ with respect to the feature *advancement*. This problem is solved by converting the multivalued feature into a vector of binary features if we use a somewhat verbose format. In general a feature with *n* values is always converted to a vector of *n-1* binary values. We illustrate this with the feature *advancement* which will be represented by three binary features:

	advance 1	**advance 2**	**advance 3**
front	1	0	0
central	1	1	0
back	1	1	1

We also need to pay special attention to affricates. When we find for example a [ʦ], we use the average values of the binary feature representations of the [t] and the [s].

As mentioned above, many consonants have palatalized counterparts. We represented e.g. [kʲ] by averaging the place of articulation of the [k] with palatal. Averaging is again done on the basis of the binary representations.

4.1.3 *Levenshtein distance*

Using the Levenshtein distance, two varieties are compared by comparing the pronunciation of words in the first variety with the pronunciation of the same words in the second. We determine how one pronunciation might be transformed into the other by inserting, deleting or substituting sounds. Costs are assigned to these three operations. In the simplest form of the algorithm, all operations have the same cost, e.g., 1. We illustrate this with an example of two varieties of a word pronunciation in northwestern dialects.

Table 2. The vowel features of Almeida & Braun and their possible values

Feature	Value	Meaning
advancement	1	front
	2	central
	3	back
height	1	close
	2	near-close
	3	close-mid
	4	central
	5	open-mid
	6	near-open
	7	open
roundedness	0	no
	1	yes

Table 3. The consonant features of Almeida & Braun represented in a binary system

Feature	Value	Meaning
place	1	bilabial
	2	labiodental
	3	dental
	4	alveolar
	5	postalveolar
	6	retroflex
	7	palatal
	8	velar
	9	uvular
	10	pharyngeal
	11	glottal
manner	1	plosive
	2	nasal
	3	trill
	4	tap or flap
	5	fricative
	6	lateral fricative
	7	approximant
	8	lateral approximant
voice	0	no
	1	Yes

Changing one pronunciation into the other can be done as follows (ignoring suprasegmentals and diacritics):

tsrɛʃna	substitute ts by ʧ	1
ʧrɛʃna	insert ɛ	1
ʧɛrɛʃna	delete n	1
ʧɛrɛʃa		
		3

In fact many sequence operations map [tsrɛʃna] to [ʧɛrɛʃa]. The power of the Levenshtein algorithm is that it always finds the cost of the cheapest mapping. Levenshtein distance is then the distance assigned by the Levenshtein algorithm, the cost of the least expensive means of mapping one string to another.

To deal with syllabicity, the Levenshtein algorithm is adapted so that only vowels may match with vowels, and consonants with consonants, with several special exceptions: [j] and [w] may match with vowels, [i] and [u] with consonants, and central vowels (in our research only the schwa) with sonorants. So the [i], [u], [j] and [w] align with anything, but otherwise vowels align with vowels and consonants with consonants. In this way unlikely matches (e.g., a [p] with an [a]) are prevented. In our example we thus have the following alignment:

ts	0	r	ɛ	ʃ	n	a
ʧ	ɛ	r	ɛ	ʃ	0	a
1	1				1	

In earlier work we divided the sum of the operation costs by the length of the alignment. This normalizes scores so that longer words do not count more heavily than shorter ones, reflecting the status of words as linguistic units. However, Heeringa, Kleiweg, Gooskens & Nerbonne (2006) showed that results based on raw Levenshtein distances approximate dialect differences as perceived by the dialect speakers better than results based on normalized Levenshtein distances. Therefore we do not normalize the Levenshtein distances in this paper.

Here we use Levenshtein as demonstrated in the examples above, i.e. with binary operation costs. One might expect the use of gradual costs to be more obvious, but in a validation study Heeringa (2004) showed that, generally speaking, the use of binary costs outperforms the use of gradual costs.

Again we need to pay some special attention to affricates and palatalized consonants. Affricates as processed as sequences of two consonants. For example the [ʦ] is processed as a [t] followed by an [s]. Following our procedure for the phone frequency method, we considered a palatal sound and its non-palatal counterpart

as fully different. For example the [k] and the [kʲ] are considered as different as the [k] and [v].

The distance between two varieties is calculated as the average of the 36 Levenshtein distances which correspond with 36 word pairs.

4.2 Design

Trubetzkoy (1930) suggested that superficial similarity in pronunciation – in the absence of regular sound correspondence – should constitute evidence of a *Sprachbund*, using Bulgarian's relation to the other Balkan languages as an example.[5]

If we add to this the conjecture that such groups originate in language contact, and that such contact is most intense near borders, then we should expect to see that pronunciation similarity is most extreme near borders, a hypothesis which we can test readily using a regression analysis, once we have settled on a suitable measure of pronunciation similarity.

We therefore measure, for each of the varieties in the subset of 50 sites, taken from the sample of 488 sites (see Section 3.2) the distance to the nearest border for each of the four contact languages. We hypothesize that the distance to the border will correlate positively with the pronunciation distance as measured by PFM, the FFM and Levenshtein distance.

5. Results

We first examine the overall measurements in order to determine which of the measurement techniques appears to be successful in detecting linguistic affinity. We then turn to the correlation with geography.

---

5. From Trubetzkoy's (1930) brief note:

Gruppen, bestehend aus Sprachen, die [...] manchmal auch äussere Ähnlichkeit im Bestande der Lautsysteme, – dabei aber keine systematischen Lautentsprechungen, keine Übereinstimmungen in der lautlichen Gestalt der morphologischen Elemente und keine gemeinsamen Elementarwörter besitzen, – solche Sprachgruppen nennen wir Sprachbünde. So gehört z.B. das Bulgarische einerseits zur slawischen Sprachfamilie [...] andererseits zum balkanischen Sprachbund [...].

Translated in English:

Groups consisting of languages which […] often have external similarities in the inventories of sound systems, but no systematic sound correspondences, no similarities in the sound shape of morphological elements and no shared elementary words, such linguistic groups we call Sprachbünde. For example Bulgarian belongs to the Slavic language family on the one hand […] and to the Balkan Sprachbund on the other hand […].

5.1 Distances to standard languages

We examine the overall measurements in two respects to see whether they were sensitive to the sort of linguistic similarity we wish to detect. First, we examined what Nerbonne & Kleiweg (2007) call *local incoherence* to see how well the measurement was detecting a signal of geographic coherence. Levenshtein distance was far and away the best technique in this respect. We applied Levenshtein distance irrespective of whether words of a comparison pair are cognates or not.

Second we checked whether the consensus view, i.e. that Macedonian is most similar to Bulgarian, followed closely by Serbian, is in fact reflected by all of the measurement techniques. In order not to be confused by the similarity of some varieties, even in the face of substantial overall differences, we examine not only the average degree of similarity, but also the degree of similarity of the most similar varieties (the first quartile of measurements).

The table shows the mean and the standard deviation of the distances to each standard language (in the two columns on the right), while the "first quartile" columns show mean and standard deviations for the closest quarter of the dialects (per language). The Levenshtein distances are averaged over the number of dialects, and over the number of words per dialect.

Table 4. The Levenshtein distances between all of the 488 Bulgarian dialects and each of the putative sources of contact influence

	first quartile		**all distances**	
	mean	**sd**	**mean**	**sd**
Macedonian	1.1	1.3	1.3	1.7
Serbian	2.6	7.0	2.5	6.1
Romanian	4.9	23.7	4.8	23.5
Greek	5.3	28.2	5.5	29.8
Turkish	5.4	29.3	5.4	29.6

To do this, we calculated the average linguistic distance between each of the reference languages and all of the 488 Bulgarian dialects for which we had data. The same descriptive statistics were calculated while restricting attention to the top 25% of most similar varieties. The results in Table 4 show the mean and the standard deviation for each standard language using the Levenshtein distance, which has been averaged over the number of dialects, and over the number of words per dialect. Levenshtein distances conform to the expectation that Macedonian is closest, followed by Serbian. Both have relatively small standard deviations. Romanian is more distant, followed by Turkish and Greek, and the distances to these standard languages have relatively high standard deviations.

It turns out that the order-insensitive methods, PFM and FFM, are only marginally less successful in detecting linguistic similarity. However, when attention is restricted to the most similar quartile, PFM and FFM agree with Levenshtein in showing that Macedonian is closest, followed by Serbian. FFM differed from the other two when the entire set of Bulgarian varieties was examined, where it led to results in which Serbian was most similar. As we noted above, the analysis of the most similar varieties is probably the better pole of comparison when examining these results.

So it turns out that PFM and FFM, which we suspected would be more suitable for the comparison of (strongly) unrelated varieties, are not clearly better. On the other hand, we do not conclude that they are clearly worse either, only marginally so. In particular, all three methods result in analyses of the first quartile of data in which the consensus view of experts is respected.

5.2 Geographic gradient of contact

We turn then to our second topic, the degree to which we can detect a gradient of similarity approaching the borders of other languages areas. We shall continue to examine alternative measurement techniques since we do not regard any as clearly superior, even if Levenshtein distance seems (marginally) preferable to the alternatives.

Figure 3 displays Levenshtein distances of 488 Bulgarian varieties compared to Macedonian and Serbian, in Figure 4 the same varieties are compared to Romanian and Greek, and in Figure 5 they are compared to Turkish. In Figure 2 the varieties are represented by dots, which represent locations. In Figures 3, 4 and 5 not only the dots are colored, but the areas surrounding the dots as well, in order to get clearer pictures. In general (nearly) the same dialect is spoken in the direct neighborhood of a location, although there may be exceptions, especially as regards (large) cities.

The Macedonian and Serbian map clearly show a gradient of similarity toward the border. But we note again here that the gradient of similarity may not indicate language contact effects at all, but rather the pronunciational residue of a continuum in the South Slavic languages. The Romanian map, the Greek map and the Turkish map do not suggest a strong gradient of similarity toward the relevant borders, but we shall examine the gradient numerically, as well. It is striking that the Greek map shows a gradient of similarity toward the Macedonian border. Bulgarian dialects which are relatively close to Greek, are close to Macedonian as well.

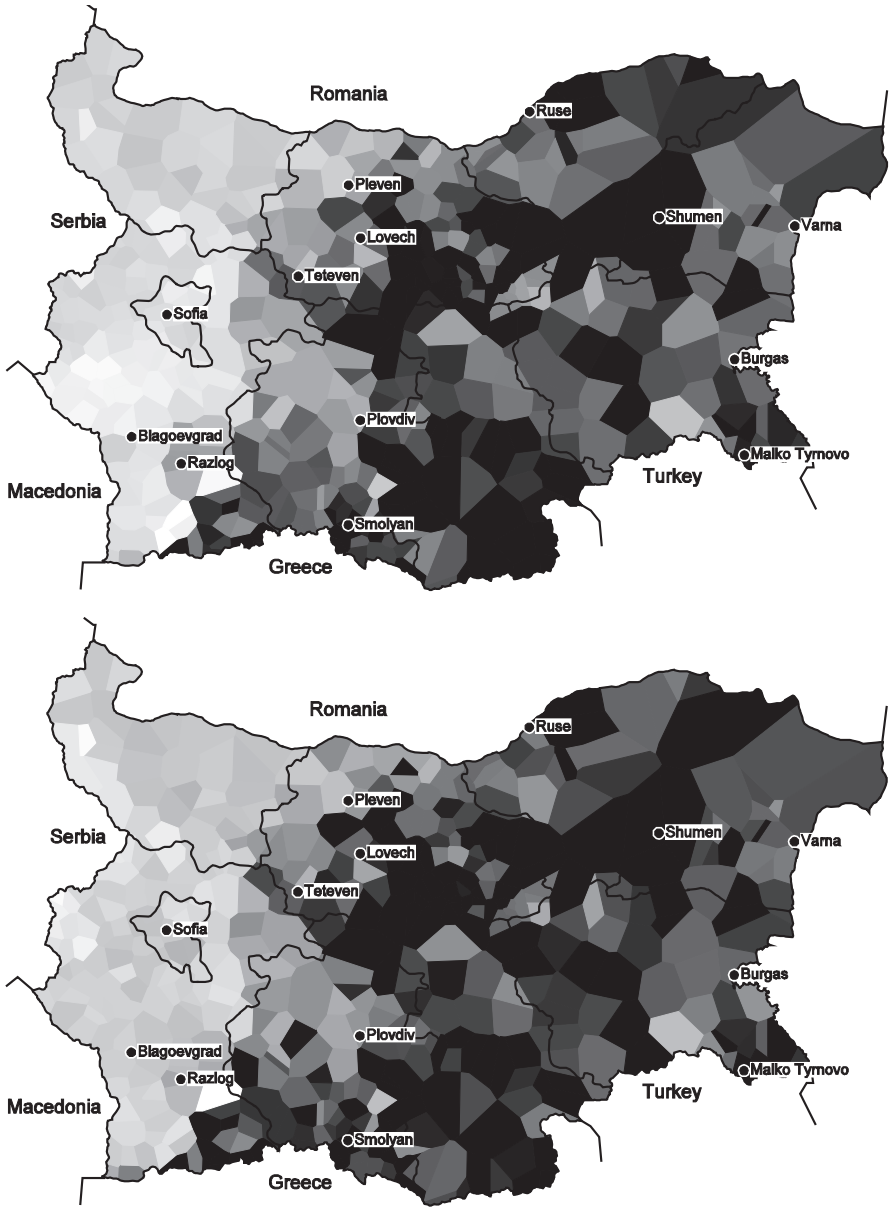

Figure 3. Average Levenshtein distances of 488 Bulgarian dialects compared to Macedonian (top) and Serbian (bottom). Dialects are represented by polygons. Lighter polygons represent closer dialects, and darker ones more distant dialects. Notice the clear gradient in similarity with respect to the western (Serbian) border.

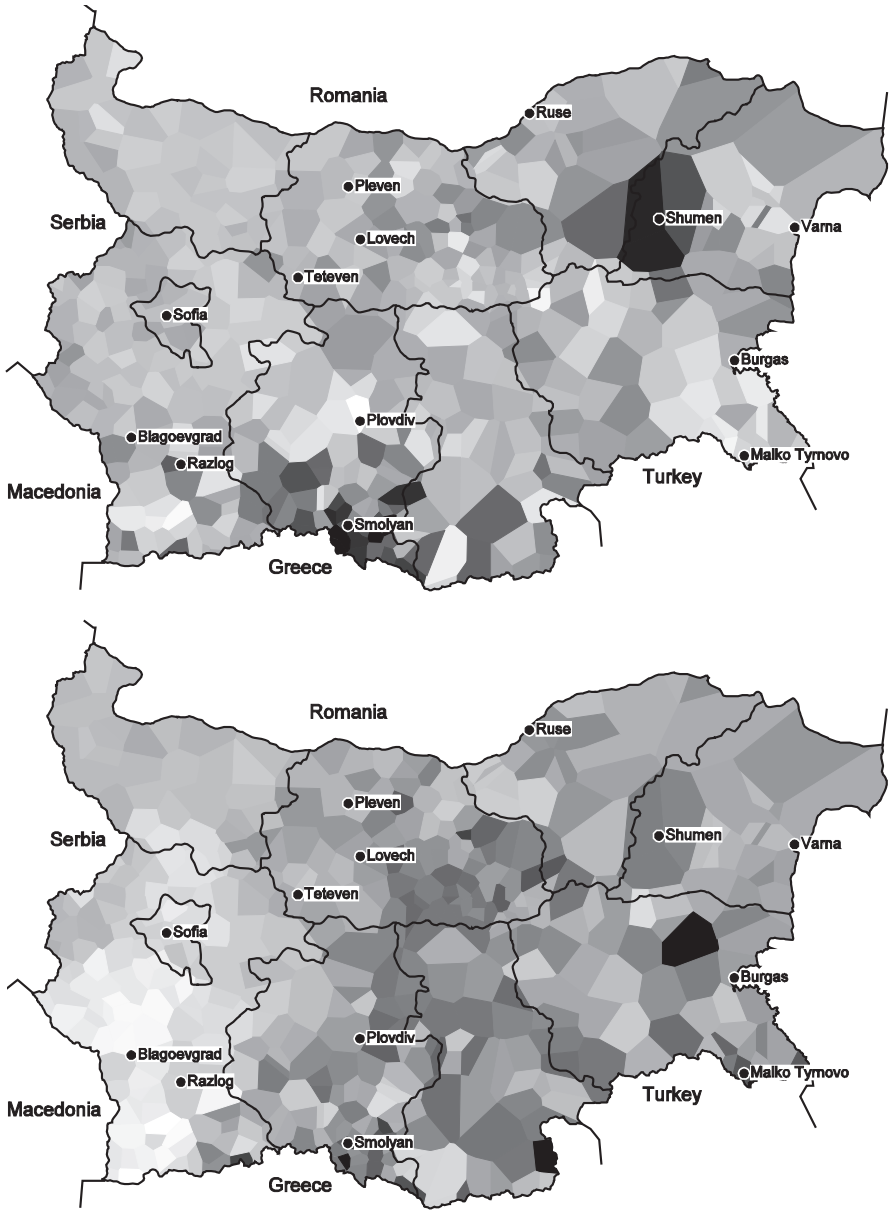

Figure 4. Average Levenshtein distances of 488 Bulgarian dialects compared to Romanian (top) and Greek (bottom). Dialects are represented by polygons. Lighter polygons represent closer dialects, and darker ones more distant dialects. We see little visual reflection of the gradients hypothesized with respect to the Romanian and Greek borders.

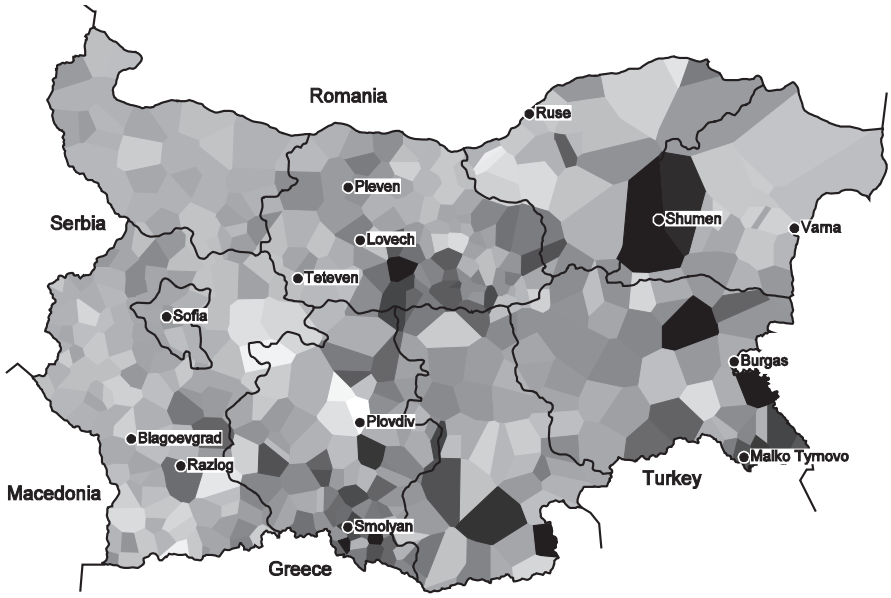

Figure 5. Average Levenshtein distances of 488 Bulgarian dialects compared to Turkish. Dialects are represented by polygons. Lighter polygons represent closer dialects, and darker ones more distant dialects. Again we see little reflection of a gradient with respect to the Turkish border.

5.3. Correlation between geographic and linguistic distances

For a subset of 50 Bulgarian dialects we measured the geographic distances to the (closest) borders of Macedonia, Serbia, Romania, Greece and Turkey. We calculated the correlations between these geographic distances and the linguistic distances to the corresponding standard languages of these countries. The results are given in Table 5. Linguistic distances were calculated using the PFM, the FFM and Levenshtein distance.

We first note that the results agree to some extent. All of the techniques detect clines of increasing similarity approaching the borders of Macedonia, Serbia and Romania, and none of them see any such (positive) gradient when approaching the Greek or Turkish border.

In fact, PFM and Levenshtein actually detect significant *negative* correlations between linguistic and geographic distances involving Greek or Turkish on the one hand and the Bulgarian varieties on the other. FFM measures a nonsignificant correlation, but again a correlation in the direction opposite from the one predicted.

Table 5. For each standard language linguistic distances to 50 Bulgarian dialects are calculated. Geographic distances are measured between the border of the corresponding country and the 50 dialects. The table shows the correlations between the linguistic distances and the geographic distances. We measured linguistic distances with the PFM, the FFM and Levenshtein distance. Correlations with $p<0.05$ are marked with *, those with $p<0.01$ are marked with ** and those with $p<0.001$ with ***.

	phone frequency method	feature frequency method	Levenshtein distance
Macedonian	0.52 ***	0.41 **	0.65 ***
Serbian	0.59 ***	0.24	0.80 ***
Romanian	0.53 ***	0.52 ***	0.34 *
Greek	−0.24	−0.56	−0.11
Turkish	−0.49 ***	−0.04	−0.21

The relatively strong correlation with Romanian when using the PFM and the FFM is all the more remarkable given the large consensus among Balkanists that pronunciation plays a subordinate role in the *Sprachbund* (Birnbaum 1965). Romanian is, of course, a Romance language, and it is surprising to see that its phonological properties are increasingly shared as one proceeds toward its borders, given the usual tenet that Balkan language contact does not involve phonology, at least not primarily. Perhaps Asenova (1989), is correct in identifying the similar vowel systems of the Balkan languages as a unifying feature (but we note that Asenova does not regard Turkish as participating in the *Sprachbund*, an issue outside the scope of this paper and one we have attempted to avoid taking a stand on). Investigating the linguistic basis of the Romanian gradient will have to await a next paper.

We return to the cases of Greek and Turkish. For Greek, both the PFM and the FFM measurements result in significant negative correlations. For Turkish the FFM measurements result in a significant negative correlation. The Bulgarian varieties we collected and analyzed become less similar to Greek/Turkish as one approaches the border. Counseled by caution, we emphasize that techniques we are applying are novel in this area so that we cannot rule out problems in the measurement techniques. But simple error is unlikely to result in statistical significance.

A more interesting conjecture for Turkish is that the explanation lies in the more complicated relation between Turkish contact and Bulgarian. After all, Bulgarian was a part of the Ottoman empire from 1393 on for nearly five centuries. Hence, the sites with substantial Turkish populations are not only located near the Turkish border, but practically all over the country. For example, there are compact Turkish populations in the Northeast (Shumen, Targovishte, Razgrad, Silistra), in the south central part of the country (Plovdiv), and in southern parts (Kardzali,

Smolyan). We would be interested in following up this conjecture with a study involving such demographics (if the relevant quantitative information is available). Linguistically we found that palatalization of [b], [t], [d], [v], [n] and [r] is most frequently found in the eastern Bulgarian dialects. However, none of these palatal sounds occur in Turkish, which makes the geographically more distant western Bulgarian varieties linguistically closer to Turkish than the eastern ones.

In the southern part of Bulgaria a large Greek population lived, especially in the area which was known as Eastern Rumelia in the period 1878–1885 when it was an autonomous province in the Ottoman empire. This population was largely exchanged in the aftermath of the Balkan wars and the second world war. Today, several thousand Bulgarians of Greek descent still inhabit the region, especially the *Sarakatsani*, transhumant shepherds. Actually we may expect a positive correlation from this, but probably the Bulgarians want to distinguish themselves from the Greek by contrasting their dialect pronunciation to the Greek pronunciation.

Although the results in Table 5 agree to some extent, the correlation measures do not agree with each other well, in particular the phone frequency and feature frequency methods as applied to Turkish and the feature frequency method and Levenshtein distance as applied to Serbian. This is important with respect to the methodological goal of developing techniques which detect contact effects. The failure of the techniques to agree indicates that they are not *all* functioning as wished.

6. Conclusions and prospects

Bulgaria and the Balkans are most famous linguistically for the extensive language contact which has developed there (Trubetzkoy 1930), and it is fascinating to apply quantitative techniques developed for dialectology in order to explore and analyze language contact.

In this paper we applied a measurement of pronunciation differences to a large database of Bulgarian.

We see the future work in several directions. First, we would like to examine different dialect data, and in particular data collected from sites that were not selected for being purely Bulgarian. Second, it would be important to identify the regular aspects of the distinctions at the base of the analysis here, i.e. the linguistic basis of the aggregate analysis, and, in fact, we have initiated that work in collaboration with a Ph.D. student. Third, it would be interesting to include lexical variation in a parallel analysis, and to examine the degree to which lexical differences correlate with differences in pronunciations. We hasten to add that a great deal more material would be needed in order to obtain reliable lexical measurements.

References

Almeida, A. & Braun, A. 1986. 'Richtig' und 'Falsch' in phonetischer Transkription; Vorschläge zum Vergleich von Transkriptionen mit Beispielen aus Deutschen Dialekten. *Zeitschrift für Dialektologie und Linguistik* LIII(2): 158–172.

Asenova, P. 1989. Балканско езикознание. *Balkan Linguistics* (in Bulgarian). Faber, Veliko Tarnovo.

Birnbaum, H. 1965. Balkanslavisch und Südslavisch. Zur Reichweite der Balkanismen im Südslavischen Sprachraum. *Zeitschrift für Balkanologie* 3: 12–63.

Chambers, J. K. & Trudgill, P. 1998, [[1]1980]. *Dialectology.* Cambridge: CUP.

Chomsky, N. A. & Halle, M. 1968. *The Sound Pattern of English.* New York NY: Harper & Row.

Heeringa, W. 2004. Measuring Dialect Pronunciation Differences Using Levenshtein Distance. PhD thesis, University of Groningen. <http://www.let.rug.nl/~heeringa/dialectology/thesis>.

Heeringa, W., Nerbonne, J., Niebaum, H., Nieuweboer, R. & Kleiweg, P. 2000. Dutch-German contact in and around Bentheim. In *Languages in Contact* [Studies in Slavic and General Linguistics 28], D. Gilbers, J. Nerbonne & J. Schaeken (eds). 145–156. Amsterdam: Rodopi.

Heeringa, W., Kleiweg, P., Gooskens, C. & Nerbonne, J. 2006. Evaluation of string distance algorithms for dialectology. In *Linguistic Distances Workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics, Sydney, July, 2006,* J. Nerbonne & E. Hinrichs (eds.), 51–62. Sydney: ACL.

Hoppenbrouwers, C. & Hoppenbrouwers, G. 1988. De featurefrequentiemethode en de classificatie van Nederlandse dialecten. *TABU* 18(2): 51–92.

Hoppenbrouwers, C. & Hoppenbrouwers, G. 2001. *De Indeling van de Nederlandse Streektalen. Dialecten van 156 Steden en Dorpen Geklasseerd Volgens de FFM.* Assen: Koninklijke Van Gorcum.

IPA. 1949. *The Principles of the International Phonetic Association: Being a Description of the International Phonetic Alphabet and the Manner of Using it, Illustrated by Texts in 51 Languages*. London: International Phonetic Association.

IPA. 2003. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet.* Cambridge: CUP.

Kurath, H. 1972. *Studies in Areal Lingusitics.* Bloomington IN: Indiana University Press.

Nerbonne, J. & Kleiweg, P. 2007. Toward a dialectological yardstick. *Quantitative Linguistics* 14(2): 148–167.

Nerbonne, J. & Kretzschmar, W. (eds). 2006. *Progress in Dialectometry*. Special issue of *Literary and Linguistic Computing* 21(4).

Osenova, P., Heeringa, W. & Nerbonne, J. Forthcoming. A quantitative analysis of Bulgarian dialect pronunciation. To appear in *Zeitschrift für Slavische Philologie*.

Stojkov, S. 2002. Българска диалектология, 4th edn. Bulgarian Academy of Science, Sofia.

Trubetzkoy, N. S. 1930. Proposition 16. Über den Sprachbund. In *Actes du premier congrès international de linguistes à la Haye du 10.–15.1928*, Vol. 1, 17–18. Leiden: A.W. Sijthoff.

Appendix

Table 1. The thirty-six Bulgarian words which formed the base of the study in phonemic transcription. The transcriptions of the Macedonian, Serbian, Romanian, Greek and Turkish equivalents are given as well. All 488 sites used in this study included phonetic transcriptions of these thirty-six words.

Bulgarian cyrillic written form	Bulgarian pronunciation	Macedonian pronunciation	Serbian pronunciation	Romanian pronunciation	Greek pronunciation	Turkish pronunciation	English translation
бъчва	ˈbətʃva	ˈbotʃva	ˈbatʃva	buˈtoj	viˈtion	ˈfɯʃɯ	barrel
зълва	ˈzəlva	ˈzolva	ˈzaova	kumˈnatə	anðaˈðɛlfi	ˈɟœɾymʤɛ	sister-in-law
дошъл	doˈʃɛl	doˈʃol	doˈʃao	vɛˈnit	ˈirəɛ	ˈɟɛlʲmiʃ	has come-he
жълт	ʒəlt	ʒolt	ʒut	ˈgalbɛn	ˈkitrinos	saˈɾɯ	yellow
зъб	zəp	zap	zup	ˈdintɛ	ˈðodi	diʃ	tooth
събота	ˈsəbota	saˈbota	ˈsubota	ˈsɨləbtə	ˈsavaton	ʤumaɾtɛsi	Saturday
къща	ˈkəʃta	ˈkukʲa	ˈkuʧʲa	ˈkasə	ˈspiti	ɛf	house
бяла	ˈbʲala	ˈbɛla	ˈbɛla	ˈalbə	aspriˈa	ak	white-fem
бели	ˈbeli	ˈbɛli	ˈbɛli	ˈalbɛ	asˈpri	ak	white-pl
язди	ˈjazdi	ˈyazdi	ˈjaʃɛ	kələɾɛʃtɛ	piɟɛˈni	aˈtabinojoʃ	ride-3per
неделя	nɛˈdɛlʲa	ˈnɛdɛla	nɛˈdɛlja	duˈminikə	kiriaˈki ɛvðoˈmaða	ˈpazaɾ	Sunday
млекар	mlɛˈkar	ˈmlɛkar	ˈmlɛkar	ləpˈtar	ɣalaktoˈpolis	ˈsyʧy	milkman
грешка	ˈgrɛʃka	ˈgrɛʃka	ˈgrɛʃka	grɛˈβʲala	ˈlaɵos	ˈhata ˈjanɯʃlɯk	mistake
венчило	vɛnˈʧilo	vɛnˈʧilo	vɛnˈʧanjɛ	kunuˈniɛ	niɦiˈkos	niˈkʲah	married life
ключ	klʲuʧ	kluʧ	kljuʧ	ˈkɛjɛ	ˈkliði	ˈanahtaɾ	key
чаша	ˈʧaʃa	ˈʧaʃa	ˈʃolja ˈʧaʃa	paˈxar	poˈtiri	ˈbaɾdak	glass; cup
път	pət	pat	put	drum	ˈðromos	jol	road
жаби	ˈʒabi	ˈʒabi	ˈʒabi	ˈbroaʃtɛ	ˈvatraçi	kuɾbaˈɣałaɾ	frogs
нощви	ˈnoʃtvi	ˈnokʲvi	ˈnatʃjvɛ	ˈkuʃkə	zumoˈtiri	ɛkˈmɛktɛknɛsi	hutch
поляна	poˈlʲana	ˈpoljana	ˈproplanak	poˈjanə	ksɛfˈto	ˈʧimɛn aˈlan	glade
овче	ˈofʧɛ	ˈofʧo	ˈofʧji ˈofʧɛtina	dɛˈoajɛ	ˈprovios	ˈkojundan	sheep's
тънко	ˈtənko	ˈtanko	ˈtanko	subˈʦirɛ	lɛpˈtos psiˈlos	ˈinʤɛ	narrow-neut
гуляй	guˈlʲaj	ˈpijanka	ˈpijanka	pɛtrɛˈʧɛrɛ kɛf	krɛˈpali ˈorjia	ʃoˈlɛn	feast
овчар	ofˈʧar	ˈovʧar ˈʧoban	ˈʧoban	ʧoˈban	vosˈkos ʦoˈpanis	ˈʧoban	shepherd
кон	kon	konʲ	konʲ	kal	ˈaloɣos ˈipos	at	horse

Bulgarian cyrillic written form	Bulgarian pronunciation	Macedonian pronunciation	Serbian pronunciation	Romanian pronunciation	Greek pronunciation	Turkish pronunciation	English translation
сън	sən	son	sanʲ	vis	ʻoniron ʻipnos	ɾyʼja	dream
отишъл	otiʲʃɛl	ʻotiʃol	otiʲʃao	plɛʼkat	piʲjɛ	ʻɟitmiʃ	has gone-he
вътре	ʻvətrɛ	ʻvnatrɛ	ʻunutra	ɨnəʼunutru	ʻmɛsa	iʼʧardɛ	inside
тенджера	ʻtɛnʤɛra	ʻtɛnʣɛrɛ	ʻlonaʦ ʻʃɛrpa	ʻkratiʦə	tɛnʣɛʼrɛs ʻtɛnʣɛɾis	ʻtɛnʤɛɾɛ	pot
джоб	ʤop	ʤɛp	ʤɛp	buzuʼnar	ʻʦɛpi	ʣɛp	pocket
няма	ʻnʲama	ʻnɛma	ʻnɛma	nuʼɛstə	ðɛniʼparçi	jok	there is no
череша	ʧɛʼrɛʃa	ʻʧɛrɛʃna	ʻʧrɛʃnja	ʧiʼrɛʃ ʧiʼrɛaʃə	kɛraʼsɛa kɛʼrasioni	ʻkiɾaz	cherry
гръб	grəp	grp	ʻlɛʤʲa	ʻspatɛ	ʻplati ʻraxi	sɯɾt	back
живя	ʒiʲvʲa	ʒiʲvɛ	ʻʒivio	trəʼit	ɛsizɛ	ʻjaʃadi	lived-he/ she/it/you
сол	sol	sol	so	ʻsarɛ	aʼlati	tuz	salt
ден	dɛn	dɛn	dan	Zi	mɛra	ɟjun	day

Language contact and phonological contrast

The case of coronal affricates in Japanese loans

Jason Shaw and Rahul Balusu
New York University

We analyze two generations of Japanese speakers' productions of [tʃi] and [ti] as they occur in native Japanese words and in loan words. Analysis across speakers verifies that this contrast is neutralized in native Japanese words and preserved in loans. Analysis of generational differences reveals two distinct patterns of preservation. Generation one speakers in our study produced overlapping distributions of [tʃi] and [ti]. In contrast, generation two speakers distinguished these strings in all environments. Our data are consistent with the view that the first generation of borrowers mapped the foreign phonological contrast to an allophonic distinction in native Japanese and that the second generation of speakers promoted this weak phonetic distinction to phonemic status.

1. Ontogeny versus diachrony in contact-induced change

This paper introduces language contact data to address the relationship between stability in the grammar of the individual and changes in the language of the community. The central research question is whether language change can be reduced to the transmission of language across generations, which we will refer to as diachronic change, or whether individuals maintain the plasticity to effectively alter their grammar in response to environmental stimuli over the course of their adult lifetime, which we will refer to as ontogenetic change.

The case study reported on here examines phonological contrast in Japanese loanwords. In most cases, contrasts that do not exist in the native vocabulary are neutralized in words borrowed from English. Since Japanese lacks, for example, a contrast between [r] and [l], the English words *lighting* and *writing* are both borrowed as [ɾaitiŋgu], neutralizing the contrast. Recently, however, it has been reported that the contrast between [ti] and [tʃi], which is non-existent in the native vocabulary, is being preserved in some loans (Shibatani 1990, Itô and

Mester 1995, Bybee 2001, Smith 2006a, b). This fact raises two related questions listed in (1) and (2) below.

(1) How did the contrast come about? Was it borrowed by a generation of speakers (ontogenetic change) or emerge with the acquisition of L1 (diachronic change)?

(2) Why is it that this contrast (as opposed to [r] and [l], for example) is selectively preserved?

The field of generative phonology, which has a history of using loanword data to test grammatical analysis, offers two disparate approaches to understanding the question in (1). One approach is to understand the emergence of new phonological contrasts as a consequence of language acquisition (see Lahiri et al. 2007 for a recent exposition of this position). Children learning the ambient language may construct qualitatively different grammars from previous generations due to an influx of loanwords or direct contact with a second language. A prediction of this approach is that true phonological contrast can only be borrowed across a generation. Adult speakers, under this account, are expected to map loanwords to a set of contrasts which exist in the native language. A competing view, developed by Smith (2004, 2006a, b), claims that speakers construct a representation of loanwords that allows for special preservation of contrast without affecting the native (non-loan) grammar. Neither of these theories, however, makes specific predictions as to which contrasts will be preserved, the question put forth in (2).

A second issue is that generative theories of phonology, which focus on describing the knowledge of individuals, typically do not consider the role that social factors may play in the spread of innovative forms through a community. Language change at the level of the community is necessarily slower than ontogenetic changes in the individual. Even if the grammar maintains the plasticity to represent a novel contrast, whether or not the form will be incorporated into the language of the community depends ultimately on sociolinguistic factors. Past work on phonetic variation in Japanese has shown vowel devoicing to be conditioned by sociolinguistic variables such as age and gender (Yuen 1997, Imai 2004). Since increased affrication in [tʃi] relative to [ti] correlates with voicelessness of the following high vowel, it is reasonable to consider the role that age and gender may have on expressing the emerging ti~tʃi contrast. This study investigates two generations of speakers in a Japanese community and identifies quantitatively different responses to expressing the [ti]~[tʃi] contrast. Analysis of social factors reveals a distributional pattern consistent with a sound change in progress (Guy et al. 1986) and, when considered in conjunction with the linguistic factors that condition variation and the etymological record of individual words, leads us towards principled answers to questions (1) and (2) above.

Evidence from an item-based analysis of age-stratified acoustic data suggests that the older group (age 50–56), *generation one*, henceforth GEN-1, acquired the ti~tʃi contrast ontogenetically. For some of these speakers, the contrast is preserved in words that entered the language during adulthood but neutralized in those that entered the language during childhood. This pattern never occurs for the younger group (20–23), *generation two*, henceforth GEN-2. Rather, consideration of the overall distribution of [ti] and [tʃi] suggests that the younger generation acquired the contrast with their first language. Thus, the answer to the first question is that, at least for some GEN-1 speakers, the contrast was borrowed synchronically.

A possible answer to the second question arises from an analysis of linguistically conditioned variation in native productions of [tʃi]. For all speakers in the study, the degree of frication produced in both [ti] and [tʃi] sequences is conditioned by prosodic structure such that stronger prosodic positions (i.e. heads of feet, accented syllables) have more frication than weaker positions. For GEN-1 speakers, however, productions of [ti] and [tʃi] in loans overlap the range of frication durations conditioned by prosody in native words. That is, it appears that prosodically conditioned variation in the realization of native [tʃi] is recruited to express the contrast between [ti] and [tʃi] in loans. Thus, the new phonological distinction in loans is parasitic on an already existing (though non-contrastive) phonetic continuum (see also Shaw 2007a). We conclude by suggesting that it is the presence of this variation in the realization of the native category that allows for the preservation of contrast in loans, providing a principled answer for question (2) above.

The broader implication is that, although contact-induced phonological change is a possible ontogenetic phenomenon, it is restricted in a principled way by the grammar of the borrowing language.

Beyond the implications for phonological theory, the current work is of general interest to the study of language contact. The community investigated is not in direct regular contact with native English speakers, but, rather, is recipient to English loan forms diffused through normalized usage in society. As such, it provides a case of language contact which is insolated, to some degree, from cross-cultural contact and can provide a reasonable baseline for more complicated interactions.

The language contact environment in Japan is characterized by mass media, wide-spread Internet access, compulsory English education and a central position in the global marketplace. To the extent that globalization increases both the policy-induced and technology-induced language contact environments that have arisen in Japan elsewhere in the world, Japan may offer a window into the future of language contact in times of globalization. Although the dynamics of language

populations determine the volume of words that will be borrowed in any given contact situation, the phonological forms that they will take are best predicted by understanding the cognitive system of the individuals that generate them. The main finding, that the ontogenetic borrowing of phonological contrast is possible, but limited by the native grammar, leads to predictions regarding which contrasts are acquirable by a given monolingual community. By understanding the degree to which indirect contact can affect a grammar, we are in a better position to quantify the additional effects of direct interaction with the source language, addressing more complicated cases involving emigrant speech communities, 2nd language acquisition and bilingualism.

2. [ti] and [tʃi] in Japanese

The affrication of coronal consonants before high front vowels is a widely attested, phonetically motivated phenomenon. There is evidence that affrication in this environment may contribute to the perceptual distinctiveness of coronal consonants (Cole and Iskarous 2001). Articulatorily, affrication is often attributed to palatization, typically understood as a change in the spatial properties of the consonantal closure. Alternatively, however, affrication has also been attributed in some languages to temporal aspects of articulation. Kim (2001, 2004, 2007) has argued that this is the case in Korean, providing articulatory and acoustic evidence that the affrication of [t] before [i] is a side effect of the tongue tip moving from an alveolar closure toward the vocalic target.

Languages that contrast affricated and non-affricated stop consonants in the environment preceding high vowels, such as the English minimal pairs *tease* and *cheese*, resist the phonetic naturalness of affrication where it would obscure the phonological contrast. On the other hand, languages that do not rely on affrication for phonological contrast are free to take advantage of both the heightened perceptual cues and (presumably) reduced effort of affricated coronals before high vowels.

Phonological evidence for the neutralization of [ti] and [tʃi] comes from morphophonemic alternations and early loanword adaptations. The verbal paradigm in (3) and (4) illustrates that [t]-final roots surface as [tʃ] only when followed by an [i]-initial suffix, as in the desiderative (3b), (4b) and the polite form (3d), (4d).

(3)		*kat*	'win-ROOT'
	a.	*kat-eba*	'win-CONDITIONAL'
	b.	*katʃ-itai*	'win-DESIDERATIVE'
	c.	*kat-anai*	'win-NEGATIVE'

	d.	*katʃ-imasu*	'win-POLITE'
	e.	*kat-oo*	'win-VOLITIONAL'
(4)		*mot*	'hold-ROOT'
	a.	*mot-eba*	'hold-CONDITIONAL'
	b.	*motʃ-itai*	'hold-DESIDERATIVE'
	c.	*mot-anai*	'hold-NEGATIVE'
	d.	*motʃ-imasu*	'hold-POLITE'
	e.	*mot-oo*	'hold-VOLITIONAL'

Neutralization of [ti] and [tʃi] is further evidenced by the adaptations of loanwords borrowed in the late 19th and early 20th century. The English loans *team* (circa 1918) and *cheese*, for example, were both historically borrowed with the affricate [tʃi]. At least as early as 1990, however, it was noted that "many younger speakers have begun to pronounce forms such as *party* and other recent loans with [t] (Shibatani 1990)." Even before the advent of the internet, non-Chinese loanwords had been steadily increasing in Japan. Shibatani reports an increase in the number of loanwords in the dictionary from 1.4% in 1859 to 3.5% in 1956 to 7.8% in 1972.

In 1991, the Japanese Ministry of Education[1] issued an official declaration on the orthographic representation of loanwords. The document established orthographic conventions for representing a number of foreign contrasts, including the affrication distinction before high vowels. For convenience in exposition we will write the contrast as *chi*, corresponding to [tʃi], and the novel orthographic form *ti*, corresponding to [ti]. The Ministry of Education declaration included *tea* and *volunteer* as examples of loanwords that should be written with the contrast preserving variant, e.g. ***ti****i* and *bolan****ti****a*, respectively. It also stipulated, however, that loanwords customarily written with the contrast neutralizing orthography such as ***ch****iketto* '*e****ti****quette*', *su****ch****iimu* '*s****tea****m*' and *purasu****ch****ikku* '*plas****tic***' will continue to be written with contrast neutralizing orthography. Even for a single word, both orthographic variants persist. Table 1 shows the number of Google hits (as of the time of data collection, December, 2004) for the orthographic variants of the English loanwords *team* and *teen*. While the contrast preserving variant *ti* is preferred at nearly a 9:1 ratio for *teen*, the more established loan, *team*, is still much more frequently represented using the contrast neutralizing *chi*[2].

---

1. As reported in: 平成三年六月二十八日 [June 28, 1991] 外来語の表記 [Declaration on Foreign Words] 内閣告示第二号 [Ministry Bulletin Issue 2]

2. Throughout this paper romanized forms of Japanese words reflect their most frequent Japanese kana representations. Thus, *chiimu,* will be used for 'team' and *tiin* for 'teen'. Whether or not orthographic differences are faithfully reproduced in pronunciation is an empirical question to which we return in Section 4.0.

Table 1. The frequency of orthographic variants in Google hits (December 2004) for two Japanese loanwords

Japanese orthography	Romanization	English gloss	Google hits
チーン	chiin	'teen(ager)'	59,400
ティーン	tiin		455,000
チーム	chiimu	'team'	10,400,000
ティーム	tiimu		66,700

Although the mapping from Japanese orthography to pronunciation is often straightforward, given the variation in orthography and the potential influence of language policy, it is problematic in the case of [ti] to assess the pronunciation of speakers from the orthographic representation. For this reason, an acoustic study of the target sequences was conducted.

3. Methods

3.1 Participants

Data was collected from 13 native speakers of Japanese living in two neighboring Tokyo suburbs. Nine speakers between the age of 50 and 56, GEN-1, and four subjects between the age of 20 and 23, GEN-2, participated. Of the 13 total participants, 11 were female and 2 were male (one male in each age group). GEN-1 participants reported having studied English for 3–5 years in secondary school; GEN-2 participants were university students and had studied for 5 years in secondary school and 1–2 years at their university. However, just two speakers, one in each age group, reported even basic conversational proficiency in English, and none had spent more than one month in an English-speaking country. All subjects had been living near Tokyo for at least 20 years and were naïve as to the purpose of the experiment.

3.2 Materials

Participants were asked to read 64 Japanese words within the carrier sentence *doo iu* _____*desuka* 'what kind of____'. Of the 64 words in the list, 34 contained a coronal stop followed by a high, front vowel. The remaining 30 words were random

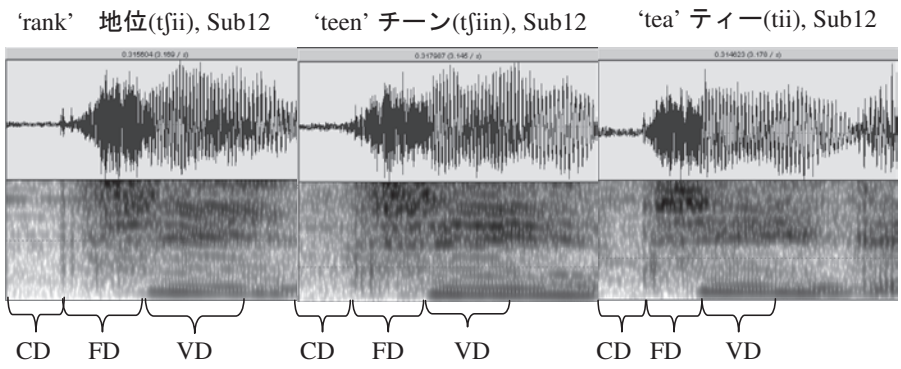

Figure 1. Sample spectrogram measurements for three tokens from one subject

fillers. Of the 34 target words (see Appendix A), 23 were loan words and 11 native[3] Japanese. The list of 64 words was randomized.

Nine of the 13 speakers read words from printed lists; three speakers from the 50–56 group read the same list from a computer screen because they were unable to read the font size on the printed sheet. Recordings were made in a quiet room using an Olympus DS-10 digital recorder and acoustic analysis was done using Praat (Boersma and Weenink 2006).

To account for orthographic variation, the use of the Japanese kana for [tɪ] and for [tʃi] were counterbalanced across tokens within each group so that each speaker's list contained half of the target words with *ti* and half with *chi* and all target words were represented equally as *ti* and *chi* within each of the two age groups. Thus, the two lists administered differed only in the orthographic representation of the target token. All subjects produced the same set of native Japanese tokens, which will serve as a control for the loan productions.

3.3 Measurements

The recordings produced 442 total data tokens (34 words * 13 subjects). Stop closure duration (CD), frication duration (FD) and vowel duration (VD) were measured for sequences of a coronal stop preceding a high, front vowel. Representative measurements are labeled in Figure 1.

---

3. For the purposes of this analysis and throughout this paper, the word "native" is used to refer to Japanese words of both Yamato and Chinese origin. Although there are certain systematic differences in the sound structure of Sino-Japanese and Yamato-Japanese words (see Itô and Mester 1995 for discussion), they both lack [ti] sequences systematically and, therefore, serve as an appropriate control condition for the study of recent loans from Indo-European languages.

The carrier sentence ensured that all target words were produced after a vowel. The CD of the target consonants was measured from the offset of voicing in the preceding vowel to the onset of the stop burst, evidenced by aperiodic energy across a broad spectrum of frequencies. The FD measure included both the stop burst and following frication period ending with the onset of voicing, as indicated by periodicity in the wave form and distinct formant structure. In some tokens the burst could clearly be distinguished from the frication period intermediate between the burst and the onset of the vowel. In other tokens it was more difficult to demarcate the burst from frication. This difficulty was compounded by the quality of the recordings, which were field quality as opposed to lab quality. For these reasons we decided to collapse the burst and following frication period into a single measure of frication duration. As a result, all of the measures included in the analysis are based upon clearly identifiable spectral landmarks.

4. Results

4.1 The language community

Analysis across speakers confirms impressionistic reports that the [ti]~[tʃi] distinction is neutralized in native words and preserved in loans, where the orthographic system has developed to represent it (see Section 2).

Figure 2. The mean measurements for CD, FD, and VD across all speakers for native [tʃi] tokens and loan [ti] tokens

Figure 2 compares the mean closure, frication, and vowel durations across speakers for *chi* (native) and *ti* (loan) words. The figure shows that, although the total duration of the CV sequence containing loan [ti] and native [tʃi] is roughly the same, there is a tradeoff between the duration contributed by the frication period and the duration contributed by the vowel. Native [tʃi] has a longer frication period and a shorter vowel; loan [ti] has a shorter frication period and a proportionally longer vowel.

We first investigated differences between native [tʃi] and loanwords represented orthographically as *chi*. Results of a one-way ANOVA on frication duration indicate no significant difference [$F(1, 295) < 1$] between these word types. We therefore collapsed native words and loanwords represented as *chi* into one group and compared this group to loanwords represented as *ti*. The results of this ANOVA were significant [$F(1, 448) = 101$, $p < .001$], indicating a reliable difference in frication duration between [ti] and [tʃi].

4.2 Results: Social factors

Having established from the pooled data that there is a significant difference in the production of [ti] and [tʃi] by the community as a whole, this section looks at the distribution of this contrast across age and gender. Our expectation is that contrast maintenance will not be uniform, but, rather, will follow the profile of a sound change in progress. Guy et al. (1986) claim that increased use of an innovating language variety by younger speakers, women, and members of lower/middle class society are predictive diagnostics of a change in progress. Since our corpus is based on the speech of a single middle-class community, we are unable to comment on the class-based distribution of the innovative form; however, we will look here at the effects of age and gender before looking more closely at the linguistic factors that condition variation in Section 5.

4.2.1 *Age*

Frication duration measurements were subjected to a two-way analysis of variance having two levels of age (GEN-1, GEN-2) and three levels of token type (native [tʃi], loan [tʃi], loan [ti]). Only the token type effect and the interaction of age and token type were significant at the $p < .05$ level. Pair-wise post-hoc comparisons found that the significant interaction is attributable to differences in the means of loan [tʃi] across generations. Figure 3 compares mean FD across generations. The significant result is indicated by an asterisk. The significant increase in FD for loan [tʃi] and the decrease in FD for loan [ti] show that GEN-2 is enhancing the contrast between [ti] and [tʃi] in loans above and beyond GEN-1.

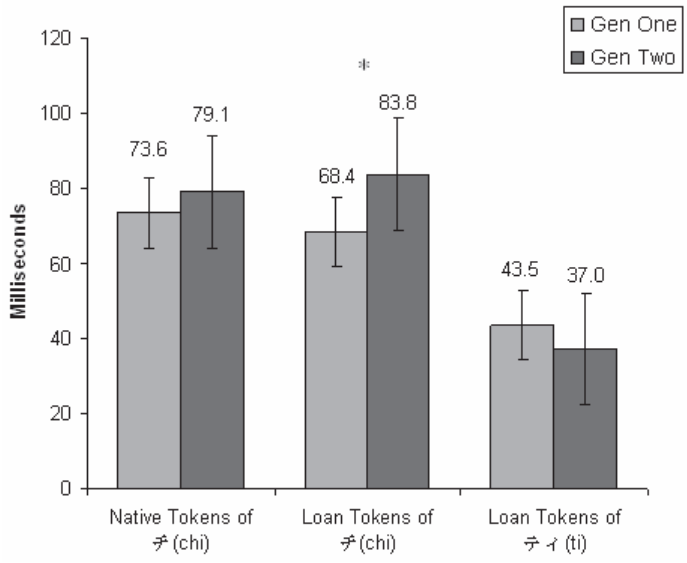

Figure 3. Comparison of frication duration across generations. Bars indicate standard error

4.2.2 *Gender*

The frication duration measurements were subject to a two-way analysis of variance having two levels of gender (male, female) and three levels of token type (native [tʃi], loan [tʃi], loan [ti]). All the effects, gender ($\underline{F}(1, 449) = 22.73$, $\underline{p} < .001$), token type ($\underline{F}(2, 449) = 4.66$, $\underline{p} < .05$) and the interaction of gender and token type ($\underline{F}(2, 449) = 7.15$, $\underline{p} < .05$) were significant at the $p < .05$ significance level.

The main effect of gender indicates that the mean frication durations were significantly different between the females (M = 61.1, SD = 33.2), and males (M = 85.3, SD = 30.8). The significant interaction, however, indicates that the gender difference in FD is not evenly distributed across the different token types. Figure 4 shows that the largest difference in FD between male and female speakers is in the production of loan tokens. Post-hoc tests show that this is the source of the interaction.

Since the gender variable picks out just two of the 13 participants in our study, it is possible that the significant effect is due to individual differences associated with these particular male speakers as opposed to an effect that is characteristic of the broader speech community. Further examination of the male subjects validates this interpretation of the effect of gender on FD. Table 2 shows that the males in the study did not perform as a homogeneous group. The GEN-2 male (age 20) produces a robust contrast between [ti] and [tʃi]. In contrast, it appears from examination of the means in Table 2 that the GEN-1 male (age 50) neutralizes the contrast.

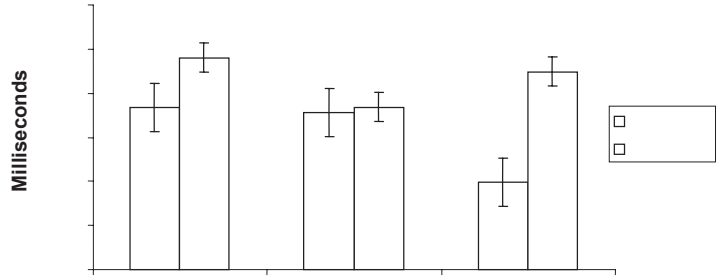

Figure 4. Comparison of frication duration across gender. Error bars indicate standard error

Since the general trend across speakers is for preservation of the [ti]~[tʃi] contrast in loans, the apparent neutralization of contrast by the GEN-1 male in the group requires explanation.

One possibility is that frication duration is a sociolinguistic variable, freely manipulated by members of this speech community. In this case, we expect the neutralizing male to be consistent across tokens (since the social context for elicitation was stable). A comparison of standard deviations between these two speakers suggests that this is likely not the case. The large standard deviation for the older male suggests an inconsistency across words that is not present for the younger male.

A second possibility is that the older speaker has acquired the contrast more recently than other speakers in the study. If the new sound [ti] can be learned without forcing a perceptual reorganization of existing categories, we would expect older words (words that came into the language earlier) to be more likely to surface as [tʃi] than words that came into the language more recently. Under this hypothesis, words represented orthographically by *ti* might fall into two phonemic categories [ti] and [tʃi]. We will test the predictions of this hypothesis by conducting a token-based analysis of this speaker in Section 6.2. First, however, it is necessary to understand the linguistic factors that might condition frication duration independently of historical or sociolinguistic factors. We take up this issue in Section 5.

Table 2. Comparison of male speakers across generation

	Loan words (ti)		**Native words**	
	Mean	**S.D.**	**Mean**	**S.D.**
Generation II (Age 20)	29 ms	(7.79)	103 ms	(20.50)
Generation I (Age 56)	94 ms	(37.2)	96 ms	(21.87)

5. Results: Linguistic factors

Results presented in Section 4.0 collapsed all tokens of a particular phoneme into a single category for analysis. It is possible that this methodology obscures sub-patterns and inconsistencies in the data. Further, if the preservation of contrast is a change in progress, as suggested by the direction of the heterogeneity with respect to age and gender, then the change is predicted to proceed more quickly in favorable linguistic environments (Guy et al. 1986). In this section, we look at variation across words and identify the linguistic factors that condition FD in [ti] and [tʃi] sequences. After identifying the factors that condition FD in native words in 5.1, we will apply the analysis to loan words in 5.2 and see the same set of conditioning factors at work. In Section 6.0, we will reconsider the effect of age in light of the linguistic factors conditioning variation.

5.1 Native words

Collapsing across all subjects, the graph in Figure 5 orders the native tokens from most to least FD in [tʃi] sequences. The two words with the greatest FD, *chii* and *chimu,* both have a pitch accent on the syllable beginning with the target phoneme. We adopt recent proposals on Japanese prosody (Yamada 1990, Haraguchi 1991, Shinohara 2000, Tuchida 2001), which analyze the language as employing left-headed (trochaic) feet built right to left with final syllable extrametricality, and the additional assumption that feet are left-aligned to morpheme boundaries (McCarthy and Prince 1993, see also Appendix A for a prosodic analysis of the stimulus set). Applying these assumptions, we see that the most fricated outputs

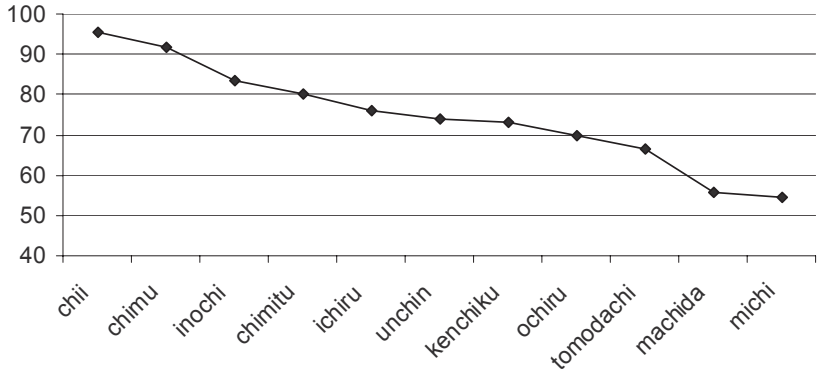

Figure 5. Native words ordered by frication duration

of [tʃi] in native words are in the strong syllable of the foot (i.e. *chimu* >> *michi*, where '>>' denotes 'more frication'). Further, Japanese words, both foreign and loan, either bear one pitch accent or are fully unaccented (Kubozono 2006). Of the words in Figure 5 in which [ti] is in the strong syllable of the foot, accented syllables have systematically longer frication than the unaccented syllables (i.e. *chimu* >> *chimitu*). Thirdly, all else being equal, syllables closer to the initial syllable in the word have more frication than subsequent syllables (i.e. *chii* >> *ichiru*). Lastly, extrametrical syllables have longer frication duration relative to other unaccented syllables (i.e. *inochi* >> *kenchiku*).

The one exception to the generalizations regarding FD reported above comes from the only verb in the word list *ochiru* 'to fall', which is not surprising since systematic differences in the prosody of nouns and verbs are common in both Japanese and cross-linguistically (see Smith 2001 for discussion). A summary of the linguistic factors conditioning FD with examples for each comparison is given in Table 3.

The above effects of linguistic structure on frication duration have analogues in studies of structurally conditioned allophony in other languages, e.g. syllable (Kahn 1976) and foot (Kiparsky 1981) level effects in English. In addition, past phonetic work on Japanese has reported durational properties that may be relevant to the effects discussed above. Although Japanese is perceived by native speakers to be isochronous with respect to the mora (Vance 1987), phonetic measurements show that the initial mora of words tends to by phonetically longer

Table 3. The left column states generalizations regarding relative frication duration between two constituents and the right two columns list specific examples of the generalization

(P_1 >> P_2), where '>>' = ' more frication'	**P_1 Examples**	**P_2 Examples**
strong syllables >> weak syllables $\sigma_s >> \sigma_w$	(**chi**i), (**chi**mu), (chimi)tu, i(**chi**ru), (**u**n)(chin), (ken)(chiku)	(tomo)(dachi), (machi)da, (michi)
accented syllables >> unaccented syllables $\sigma_a >> \sigma$	(**chi**i), (**chi**mu), i(**chi**ru)	(**u**n)(chin), (ken)(chiku), (tomo)(dachi), (machi)da, (michi)
left-edgedness >> right-edgedness $\sigma_n >> \sigma_{n+1}$	(**chi**i), (**chi**mu), >> i(**chi**ru), o(**chi**ru) (chimi)tu >> (**u**n)(chin), (ken)(chiku)	
extrametrical >> footed, unaccented <σ> >> σ	(**i**no)chi	(chimi)tu, (**u**n)(chin), (ken)(chiku), (tomo)(dachi), (machi)da, (michi)

than subsequent moras (Port et al. 1987). Similarly, analysis of extrametrical syllables has also uncovered an increase in duration over footed syllables (Teranishi 1980 as cited in Poser 1990). These effects of general syllable length are possibly related to the increase in FD in these positions.

The linguistic factors conditioning frication duration, we will argue, have important implications for the study of the emerging [ti]~[tʃi] contrast in Japanese loans. The average production of native words ranged between 95 ms of frication for *chii* and 55 ms for *michi*, with a mean of 75 ms (Figure 2). If the frication duration of loan words is subject to the same effects of prosody, given a mean frication duration of 42 ms (Figure 2), then [ti] and [tʃi] tokens across speakers will overlap considerably, leading to neutralization in some environments. Unraveling which speakers neutralize in what words will provide the key to understanding how this novel phonological contrast made its way into the language.

5.2 Loan words

Figure 6 shows the average frication duration across subjects by token for loanwords with [ti] in a strong syllable[4]. As in the native words, the duration of frication is subject to considerable gradience, spanning from 60 ms in *tii* to 30 ms in *chippu*. Further, the same set of factors used to explain the gradience in native words can go a long way towards accounting for the relative length of frication in loans. Of the words with [ti] in a strong syllable, those with the greatest frication were accented and in initial position (*tii, tiimu, chiketto, tiin*); the next most frication was found in *suchiru*, which is accented but in the second syllable, and *tiiruumu*, which is unaccented, but in the leftmost syllable of the word; the remaining words all cluster between 30–40 ms and, with three exceptions (*tiida*, *tinpanii*, *chippu*)[5], have [ti] at least two syllables from the left edge of the word.

4. There were no loanwords in the stimulus set that had [ti] in a weak syllable. All loan tokens were either in the strong syllable of a foot or extrametrical (see Appendix A for prosodic analysis of the stimulus set).

5. Although we will not pursue an analysis of the exceptions here, the following factors may be relevant: (1) *chippu* is the only word in the study that has consecutive vowels in devoicing environments. As such, the first vowel is somewhat special in that it is a voiced vowel between two voiceless consonants (c.f. *chiketto* with one devoiced vowel) and would devoice if it was not for the following devoiced vowel (Tuchida 2001). It's a reasonable possibility that the unnatural maintenance of voicing affects the frication duration of the previous consonant. (2) *tiida* and *tinpanii* are the lowest frequency words in the study (as indicated by Google hits). Although there were no general effects of frequency, it has been suggested in the literature that frequency of use may correlate with degree of nativization (Itô and Mester 1995).

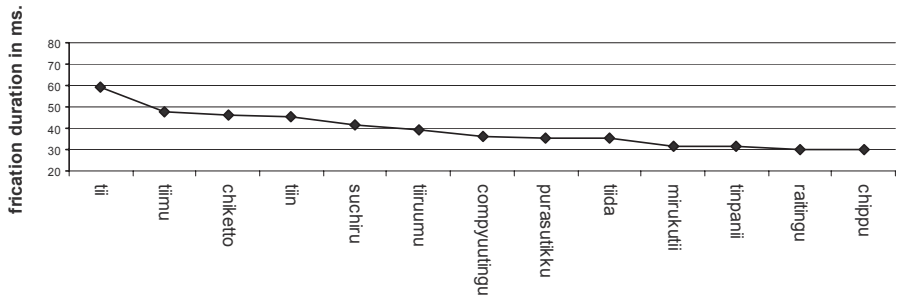

Figure 6. The frication duration of loanwords with [ti] in strong syllables

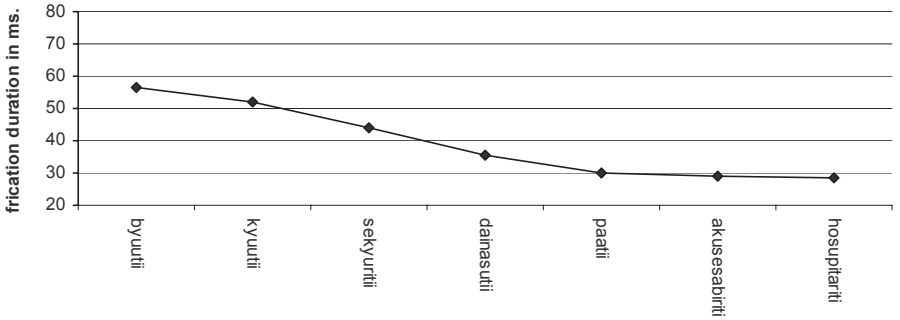

Figure 7. Frication duration of loanwords with [ti] in extrametrical syllables

Figure 7 shows the average frication duration across subjects for loanwords with [ti] in extrametrical syllables. Since they are not footed, these syllables cannot be accented. Thus, the only factor that predicts the relative frication duration is proximity to the initial syllable. The extrametrical syllables closest to the left edge of the word are in disyllabic *byuuti* and *kyuuti* and, indeed, have the greatest frication duration. These are followed by the four-syllable words *sekyuritii* and *dainasutii* which are in turn followed by even longer words *akusesabilitii* and *hosupitalitii*. The one exception to the generalization is *paatii*, which, at two syllables is predicted to be longer than *sekyuritii* and *dainasutii*.

As illustrated by Figures 5–7, where the stimulus sets allow comparison, the general trends identified for native words hold for loan words as well. FD is conditioned by prosodic position and distance from the left edge of the word. Further, the average frication duration of some [ti] words is actually longer than for some [tʃi] words. This is illustrated in Figure 8, which plots the average FD of [ti] and [tʃi] for all the words in the study against the positional strength of

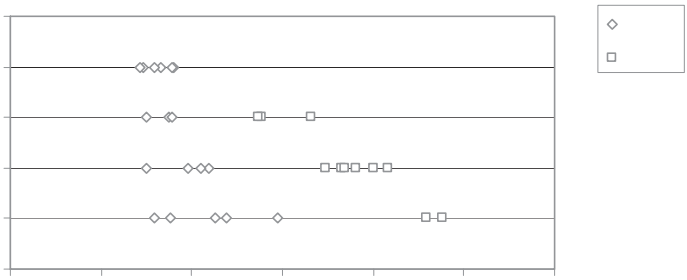

Figure 8. The average frication duration produced by all speakers for loanword [ti] and native word [tʃi] tokens plotted against the positional strength of the target sequences

the target syllable[6]. The chart shows that contrast between [ti] and [tʃi] is only achieved if the primary cue distinguishing the categories is relativized to prosodic position. At each step on the positional strength scale, there is a contrast of at least 20 ms in frication; however, if we view the categories as independent of prosodic positions, then there is distributional overlap.

In this section, we identified a number of factors that condition the realization of [ti] and [tʃi]. Speakers produce a consistent contrast between these categories when they are in analogous prosodic positions, but the complete distribution of [ti] overlaps with [tʃi]; [ti] in strong positions is produced with more FD than [tʃi] in weak positions. The observations discussed in this section, however, have been true of the language population as a whole, holding for the average FDs produced by 13 speakers from two different generations. In the next section, we return to the issue of cross-generational differences and find that understanding how prosody effects FD holds the key to addressing the ontogeny vs. diachrony issue raised at the outset.

6. Position strength is represented by whole numbers 1–4 (where 1 is the strongest position, an accented syllable in the first syllable of the word, and 4 is the weakest). See Appendix B for definitions of strength categories.

6. Age revisited: Adult plasticity and phonetic enhancement

Section 5.0 identified a number of linguistic factors that condition FD in [ti] and [tʃi] sequences. In this section we revisit the effect of age on FD in light of these results and uncover clear generational differences that were obscured by the analysis in 4.2.1 which failed to take positional strength into account.

6.1 A generation gap

In Section 4.2.1 (Figure 3), we took a first look at age as a possible conditioning factor of frication duration. Comparing [ti] tokens from loanwords produced by GEN-1 and GEN-2 yielded no significant difference. When we looked at the average frication duration by word in Section 5.0, however, we saw that some loanwords contained a [ti] sequence produced with greater FD than some sequences of [tʃi] in native words. The chart in Figure 9 isolates GEN-1 speakers in comparing the FD of loan and native tokens relativized to positional strength; Figure 10 does the same for GEN-2.

Recall that Figure 8 showed that, collapsing across the entire speech population, there was an overlap in FD between [ti] and [tʃi] sequences. Figure 9 and 10 show that although both GEN-1 (Figure 9) and GEN-2 (Figure 10) speakers maintain a contrast between [ti] and [tʃi] in each position, GEN-1 produces the sequences with overlapping FD while GEN-2 produces the sequences without overlap. Thus, the overlap between FD that we saw in the population as whole can be attributed solely to GEN-1 speakers.

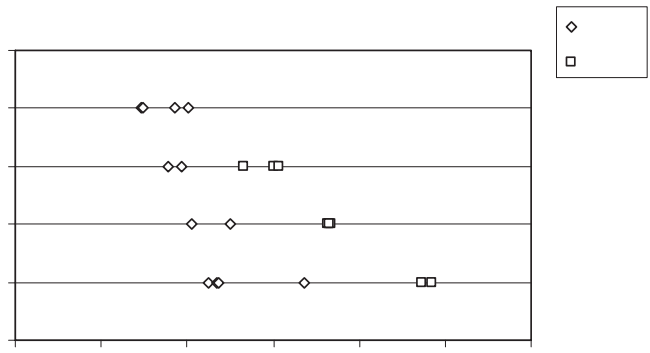

Figure 9. The average frication duration produced by GEN-1 speakers for loan [ti] and native [tʃi] tokens plotted against the positional strength of the target sequences

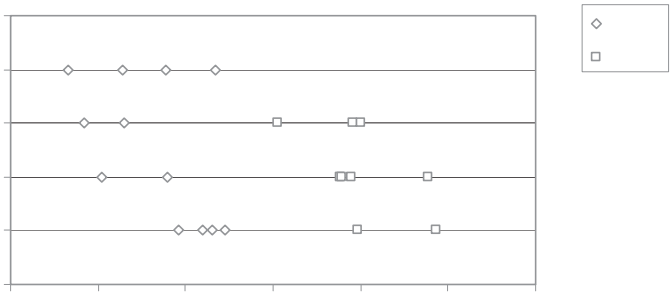

Figure 10. The average frication duration produced by GEN-2 speakers for loanword [ti] and native word [tʃi] tokens plotted against the positional strength of the target sequences location

The preservation of contrast between [ti] and [tʃi] for GEN-2 is evidenced by a sharp divide between categories at around 60 ms of frication. This contrasts with the plot of GEN-1 speakers, which has both loanwords with greater than 60 ms of frication and native words with less. From a perceptual standpoint, then, a token with less than 60 ms of frication is unambiguously [ti] only if produced by a GEN 2 speaker. If produced by a GEN-1 speaker, it could just as easily be a [tʃi] in weak prosodic position. Thus, while GEN-2 speakers show distinct categories of [ti] and [tʃi], GEN-1 speakers have overlapping categories that distinguish contrast only if relativized to prosodic position. What can account for this difference in profile? To investigate this question further, we turn to the GEN-1 subject with the most overlap between categories.

6.2 Profile of a synchronic borrower

As illustrated in Section 4.2.2 (Table 2), the male representative of the GEN-1 group produced similar mean frication durations for both native and loan stimuli. Although this fact by itself suggests that he does not have a contrast between [ti] and [tʃi], the suspiciously high standard deviation warrants a closer look. We hypothesized at the end of Section 4 that the high variance in this measure may be due to the differentiation of words orthographically expressed as *ti* into separate [ti] and [tʃi] categories. The chart in Figure 11 plots the frication duration of [ti] in all target words produced by this subject. The native words are on the left, ordered from least to greatest frication duration; the loanwords are on the right, also ordered from least to greatest frication duration. Although the slope of the line connecting points in Figure 11

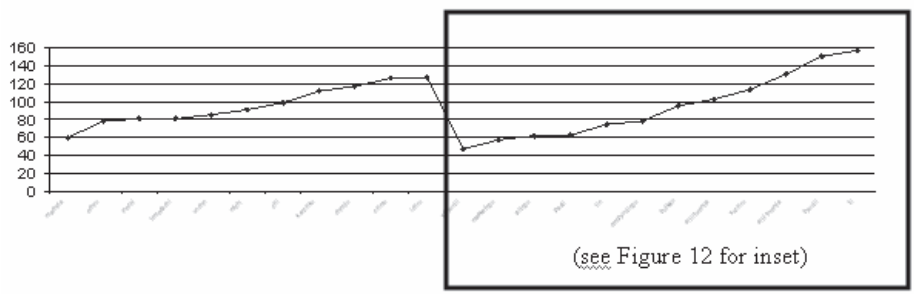

Figure 11. The frication duration of all target words produced by one GEN-1 subject. The native words are on the left ordered from least to greatest frication. The loan words are on the right also ordered from least to greatest frication duration

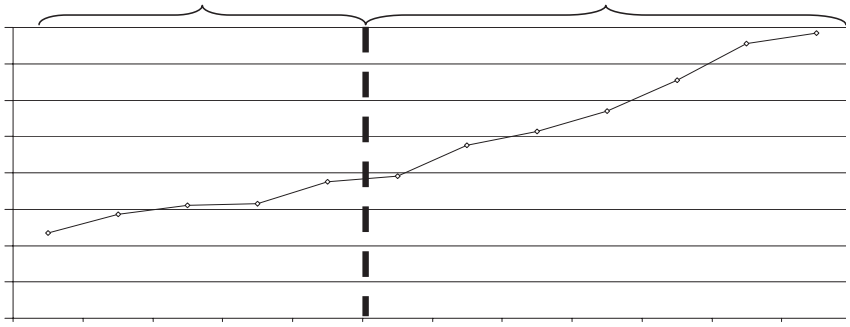

Figure 12. The frication duration of all loan words produced by the GEN-1 male subject

is similar for native and loan words, the extrema are quite different. The range of frication durations for native words runs from 127 ms in strong prosodic positions to 60 ms in weak positions. In loans, however, strong positions are produced with 157 ms of frication duration and, more interestingly, weak positions are produced with just 47 ms, which falls within the group range for [ti] (see Figure 6, 7).

To take a closer look at the tokens that are produced within the [ti] range, the graph in Figure 12 isolates the loanwords, just the right-hand portion of Figure 11. Building on the findings for the population reported in the previous section, we expect prosody to systematically condition the range of variation within a category. Figure 12 shows, however, that these expectations are born out rather

curiously. Unexpectedly, the prosodic strength hierarchy recycles in the middle of the slope. The word with the greatest frication, *tii*, is both accented and initial, placing it ahead of accented targets in the second syllable (e.g. *suchiiru*), which are in turn more fricated than the unaccented second syllable [ti]'s in *ai tii furontia*[7] and *butikku*. Both *byuutii* and the final syllable in *ai tii furontia* are extrametrical, but, as predicted, *byuutii* has a longer frication period by virtue of its proximity to the left edge of the word relative to *ai tii furontia.* This pattern accounts for only the right half of the slope in Figure 12. After *kompyuutingu* the prosodic strength hierarchy repeats. That is, *tiin* has [ti] in a strong prosodic position just as *tii* did. Both accented and word-initial, the [ti] in *tiin* is in the strongest possible prosodic position. Further, the next two words, *paatii* and *raitingu*, are in the number two slot in the strength hierarchy. The [ti] in *paatii* is extrametrical, but, like *byuutii*, being just one syllable from the left edge, it is in the strongest possible extrametrical syllable. Similarly, *raitingu*, with an accented second syllable [ti] is just one step down on the strength hierarchy from *tiin*. The words with the shortest frication, *maaketingu* and *sekyuriti* are at the bottom of the strength hierarchy. We saw for native words that FD is predictable from the prosodic strength of the syllable. This generalization carries over straight-forwardly to this speaker only if we posit two sound categories for loan sequences orthographically represented as [ti]. The proposed category boundary is marked with a dotted line in Figure 12.

Thus, if we consider the linguistic factors that condition gradience within a category, the data strongly suggest that, for this speaker, a sound treated uniformly in the source language maps to two separate categories. The *ti* in *tii* maps to a different category than the *ti* in *tiin*. Figure 11 shows, however, that both of these loan *ti*'s overlap heavily (though not completely) with native productions.

The difference between the GEN-1 speaker data in Figure 12 and the general trend for GEN-1 speakers, therefore, is not neutralization versus preservation. Rather, it appears that all speakers from GEN-1 acquired the contrast in at least some words. The speaker in Figure 12, however, treats two sets of *ti* tokens in loans as categorically different.

Why should the contrast be confined to certain words? This fact is compatible with two explanations: (1) the speaker learned the contrast during L1 acquisition, but simply misclassified some [ti]-tokens as [tʃi] or (2) the speaker acquired an L1 without a [ti]~[tʃi] contrast, but learned the contrast later in life. Evidence that the latter of these possibilities is correct can be mustered by looking at the etymological record of the loanwords produced as [ti].

7. The word *ai tii furontia* is listed twice because it has two target sequences, one in the second syllable and one in the final syllable. The second syllable has longer frication duration than the final syllable.

Of the five words produced with the least frication, *tiin, paatii, raitingu, maaketingu* and *sekyuritii*, only *paatii* was in the language when the speaker was born[8]. According to the unabridged Japanese dictionary (Nihon Kokugo Daijiten 1970–1976, 2000–2001), the next word to come into the language was *maaketingu*, in 1975, well into the adulthood of all GEN-1 speakers in the study. Thus, although the contrast is restricted to loanwords in even the second generation speakers, for this particular speaker, the contrast is further restricted to recent loanwords or, more precisely, those that entered the language after 1975.

There appear to be two effects of language contact on this speaker's grammar. Using the effect of prosodic position on FD as a diagnostic for sound category boundaries, we identified a partition in the speaker's production of *ti* into two categories. In one category, the maximum frication duration exceeds that of the native category in analogous prosodic position; in the other, the FD is far less than the native category in comparable positions. The FD in this second category comes closest to approximating the source language and is realized in words that came into the language well after the speaker's critical period. Further, the FD separating these categories is of a magnitude that is used for contrastive purposes elsewhere in the language (e.g. the maintenance of the [ti]~[tʃi] contrast before back vowels as in [ta] 'rice paddy' and [tʃa] 'tea'). Thus, the first consequence of language contact is the emergence of a new sound sequence, [ti], which contrasts minimally with the native sequence [tʃi]. A second consequence is the phonetic drift within the loan categories. The loanwords produced as [tʃi] have even greater frication than native [tʃi] such that, as shown in Figure 12, the native word [tʃii] 'rank' is produced with 100 ms of frication while the loanword [tii] 'the letter T' is produced with 157 ms, different still from [tiin] 'teen(ager)', produced with 76 ms. Thus, in addition to the formation of a new sound category, the loan categories of [ti] and [tʃi] also appear to be drifting away from each other. This reflects a larger cross-generational trend. Recall that even when collapsing across prosodic positions GEN-2 produced loan [tʃi] with greater FD and loan [ti] with less FD than GEN-1 (see Figure 3).

Taken together the two effects of language contact apparent in this speaker's productions provide insight into the process of phonological contrast borrowing. First, GEN-1 produces a weak contrast between [ti] and [tʃi] in loan words by mapping these sequences to prosodically conditioned allophones of native [tʃi] (see Shaw 2007b for formal implementation). Second, once contrasting categories are formed, the phonetic realization of the categories is slowly pushed apart. This second effect can be observed both on a small scale within individuals (Figure 11, 12) and on a larger scale across generations (Figure 3, 9, 10).

8. The speaker was 56 years old when data was collected in 2004.

7. Conclusion

At the outset of this paper, we raised two questions regarding the emergence of the phonological contrast between [ti] and [tʃi] in Japanese. The first is whether the contrast emerged within the grammar of the individual or across a generation. The evidence presented in Section 6.2 suggests that the contrast was borrowed initially by GEN-1 speakers, but was in large part parasitic on the allophonic variation used to express the native sequence [tʃi] in different prosodic positions (see Shaw 2007a for further discussion). The weak contrast produced by GEN-1 speakers was subsequently enhanced by GEN-2 speakers. The second question was why the ti~tʃi contrast was selectively preserved. Contrast preservation in loanword adaptation is not normal. As Table 4 shows, there are numerous cases in Japanese in which phonological contrasts from the source language are neutralized. In response to this question, we hypothesized that the presence of a non-contrastive phonetic continuum (i.e. allophonic variation) may be prerequisite for borrowing phonological contrast. By this hypothesis, the neutralization of contrasts in Table 4 is due to a lack of variation in corresponding native sequences.

Table 4. Distinct English forms which map to the same phonetic form in Japanese

English word		**Japanese adaptation**	
orthography	**IPA**	**orthography**	**IPA**
a. writing	ɹaiɾiŋ	ライティング	ɾaitingu
b. lighting	laiɾiŋ		
c. steal	stil	スチール	sutʃiiɾu
d. still	stɪl		
e. food	fud	フード	ɸuudo
f. hood	hʌd		

Both of these conclusions stem from an analysis of phonetic data that considers the phonological structure of the target words, the age of the speakers and the etymological record of the loans. Although it was not apparent in our first look at age as a factor conditioning FD (see 4.2.1), looking at the effect of age within prosodic position revealed clear differences between the phonetic profiles of GEN-1 and GEN-2 populations. GEN-1 speakers produce [ti] and [tʃi] with overlapping FD distributions and with contrast maintenance relativized to prosodic position. For these speakers, loan [ti] in prosodically strong positions is produced with the same temporal properties as native [tʃi] in prosodically weak positions. GEN-2 speakers, however, produce distinct FD's for [ti] and [tʃi] regardless of prosodic position. Our evidence for ontogenetic borrowing came from an analysis of a

GEN-1 speaker with lexically specific preservation of [ti]. Understanding the effect of prosodic position on FD allowed us to first identify a category boundary between loan [ti] and loan [tʃi] and subsequently recognize the contrast preserving [ti] productions as loans that entered the language during adulthood.

In sum, we have produced phonetic evidence that phonological contrasts can be borrowed and that they can be borrowed by mature adult speakers even without substantial direct contact with the source language. The evidence presented suggests that the necessary prerequisite is a phonetic continuum to which the non-native contrast can be mapped. As such, the range of contrasts which can be incorporated via this type of ontogenetic borrowing may be limited to the range of allophonic variation present in the borrowing language.

References

Boersma, P. & Weenink, D. 2006. Praat: Doing phonetics by computer, version 4.3.14. <http://www.praat.org>.

Bybee, J. 2001. *Phonology and Language Use*. Cambridge: CUP.

Cole, J. & Iskarous, K. 2001. Effects of vowel context on consonant place identification: Implications for a theory of phonologization. In *The Role of Speech Perception in Phonology*, E. Hume & K. Johnson (eds), 103–122. San Diego CA: Academic Press.

Guy, G., Horvath, B., Vonwiller, J., Daisley, E. & Rogers, I. 1986. An intonational change in progress in Australian English. *Language in Society* 15: 23–52.

Haraguchi, S. 1991. *A Theory of Stress and Accent*. Dordrecht: Foris.

Imai, T. 2004. Vowel Devoicing in Tokyo Japanese: A Variationist Approach. PhD dissertation, Michigan State University.

Itô, J. & Mester, A. 1995. Japanese phonology. In *The Handbook of Phonology*, J. Goldsmith (ed.), 817–847. Oxford: Blackwell.

Kahn, D. 1976. Syllable-based Generalizations in English Phonology. PhD dissertation, MIT. (Distributed by Indiana University Linguistics Club. Published 1980, New York NY: Garland).

Kim, H. 2001. A phonetically based account of phonological stop assibilation. *Phonology* 18: 81–108.

Kim, H. 2004. Stroboscopic-Cine MRI data on Korean coronal plosives and affricates: Implications for their place of articulation as alvealor. *Phonetica* 61: 234–251.

Kim, H. 2007. Stroboscopic-cine MRI and Acoustic data on gradual tongue movements in Korean Palatization: Implications for its coarticulatory effect. *Proceedings of ICPhS XVI*. Saabrücken, Germany.

Kiparsky, P. 1979. Metrical structure assignment is cyclic. *Linguistic Inquiry* 10: 421–41.

Kubozono, H. 2006. Where does loanword prosody come from? A case study of Japanese loanword accent. *Lingua* 116: 1140–1170.

Lahiri, A., Allison W. & Jönsson-Steiner, E. 2007. Tones and loans in the history of Scandinavian. In *Tones and Tunes: Typological Studies in Word and Sentence Prosody*, C. Gussenhoven & T. Riad (eds.), 353–376. Berlin: Mouton.

McCarthy, J. & Prince, A. 1993. Generalized alignment. *Yearbook of Morphology*, 79–153.

Nihon Kokugo Daijiten. 1970–76. (20 vols.). Tokyo: Shogakukan. (*The Encyclopedic Dictionary of the Japanese Language*, 1st edn).

Nihon Kokugo Daijiten. (2000–2001). (14 vols.). Tokyo: Shogakukan. (*The Encyclopedic Dictionary of the Japanese Language*, 2nd edn).

Port, R., Dalby J. & O'Dell, M. 1987. Evidence for mora timing in Japanese. *Journal of the Acoustical Society of America* 81: 1574–1585.

Poser, W. J. 1990. Evidence for foot structure in Japanese. *Language* 66: 78–105.

Shaw, J. 2007a. /ti/~/tʃi/ contrast preservation in Japanese loans is parasitic on segmental cues to prosodic structure. *Proceedings of ICPhS XVI*. Saarbrücken, Germany.

Shaw, J. 2007b. Structure mapping in loan phonology. Handout from *Experimental Approaches to Optimality Theory*, May 18, Ann Arbor, Michigan.

Shibatani, M. 1990. Japanese. In *The World's Major Languages*, B. Comrie (ed.), Cambridge: CUP.

Shinohara, S. 2000. Default accentuation and foot structure in Japanese: Evidence from Japanese adaptations of French words. *Journal of East Asian Linguistics* 9(1): 55–96.

Smith, J. 2001. Lexical category and phonological contrast. In *Papers in Experimental and Theoretical Linguistics* 6: *Workshop on the Lexicon in Phonetics and Phonology*, R. Kirchner, J. Pater & W. Wikely (eds), 61–72. Edmonton AB: University of Alberta.

Smith, J. 2004. Loan phonology is not all perception: Evidence from Japanese loan doublets. In *Japanese/Korean Linguistics* 14, T. Vance & K. Jones (eds), 63–74. Stanford CA: CSLI.

Smith, J. 2006a. Correspondence theory vs. cyclic OT: Beyond morphological derivation. In *Proceedings of NELS 36*, C. Davis, A. Deal, & Y. Zabbal (eds). Amherst MA: GLSA.

Smith, J. 2006b. Source similarity in loanword adaptation: Correspondence theory and the posited source-language representation. In *Phonological Argumentation: Essays on Evidence and Motivation*, S. Parker (ed.). London: Equinox.

Teranishi, R. 1980. Two-mora-cluster as a rhythm unit in spoken Japanese sentence or verse. Text of talk abstracted in *Journal of the Acoustical Society of America 67*, Supplement 1:S40

Tuchida, A. 2001. Japanese vowel devoicing: cases of consecutive devoicing environments. *Journal of East Asian Linguistics* 10(2): 225–245.

Vance, T. 1987. *An Introduction to Japanese Phonology*. Albany NY: State University of New York Press.

Yamada, E. 1990. Stress assignment in Tokyo Japanese (1) and (2). In *Fukuoka Daigaku Jinbun Ronsoo* 21: 1575–604 & 22: 97–154. Fukuoka University.

Yuen, C. 1997. Vowel Devoicing and Gender in Japanese. MA thesis, UCSD.

Appendix A

Stimulus list

Bold syllable denotes the presence of a pitch accent; Romanization of loanwords below is based on the most frequent orthographic variant; *italicized* vowels are in devoiced environments.

Japanese (Romanization)	Assumed footing	Gloss	Type	[ti] syllable strength
地位 (**chi**i)	(**tʃi**i)	'rank'	Native	1
痴夢 (**chi**mu)*	(**tʃi**mu)	'foolish dream'	Native	1
命 (**i**nochi)	(**i**no)chi	'life'	Native	2
ち密 (chimitu)	(tʃimi)tu	'precision'	Native	2
一縷 (**ichi**ru)	i(**tʃi**ru)	'gleam' (i.e. gleam of hope)	Native	2
運賃 (**u**nchin)	(**u**n)(tʃin)	'freight'	Native	2
建築 (kenchiku)	(ken)(tʃiku)	'architecture'	Native	2
落ちる (**ochi**ru)	o(**tʃi**ru)	'to fall'	Native	2
友達 (tomodachi)	(tomo)(datʃi)	'friend'	Native	3
町田 (machida)	(matʃi)da	a surname	Native	3
道 (michi)	(mitʃi)	'road'	Native	3
ティー (**ti**i)	(**ti**i)	'tea'	Loan	1
チケット (**ch*i***ketto)	(**tʃ*i***)(ket)to	'ticket'	Loan	1
スチール (s*u***chi**ru)	su(**tʃi**ru)	'steal', 'still'	Loan	2
チーム (**ti**imu)	(**ti**i)mu	'team'	Loan	1
ミルクティー (mirukuti**i**)	miruku(**ti**i)	'milk tea'	Loan	4
アクセサビリティ (akusesa**bi**riti)	akusesa(**bi**ri)ti	'accessibility'	Loan	4
ティーダ (**ti**ida)	(**ti**i)da	'Tiida' (Japanese car)	Loan	1
ビューティー (**byu**utii)	(**byu**u)tii	'beauty'	Loan	1
キューティー (**kyu**utii)	(**kyu**u)tii	'cutie'	Loan	1
ホスピタリティー (hosupi**ta**riti)	hosupi(**ta**ri)ti	'hospitality'	Loan	4
セキュリティー (se**kyu**ritii)	se(**kyu**ri)tii	'security'	Loan	2
ティンパニー (**ti**npanii)	(**tin**)panii	'timpani'	Loan	1

* novel compound

Japanese (Romanization)	**Assumed footing**	**Gloss**	**Type**	**[ti] syllable strength**
IT フロンティア (ai tii fu**ro**ntia)	(ai) (tii) fu(**ro**n)tia	'IT frontier'	Loan	2
ティールーム (tii**ru**umu)	tii(**ru**u)mu	'tea room'	Loan	2
チップ (**chip**pu)	(**tʃip**)pu	'tip', 'chip'	Loan	1
コンピューティング (kompyuu**ti**ngu)	kompyuu(**tin**)gu	'computing'	Loan	4
ティーン (**ti**in)	(**ti**in)	'teen(ager)'	Loan	1
ダイナスティ (dai**na**suti)	dai(**na**su)ti	'Dynasty' (TV program)	Loan	3
ライティング (rai**ti**ngu)	rai(**tin**)gu	'writing'	Loan	3
パーティー (**pa**atii)	(**pa**a)tii	'party'	Loan	2
マーケティング (maa**ke**tingu)	maa(**ke**)tingu	'marketing'	Loan	4
ブティック (**bu**tikku)	(**bu**)tikku	'boutique'	Loan	3
プラスチック (purasu**chi**kku)	purasu(**tʃik**)ku	'plastic'	Loan	4

Appendix B

Prosodic strength categories

Note: moras are counted from the left morpheme boundary

1 = extrametrical syllable in the 3rd mora or accented syllable in the 1st mora
2 = extrametrical syllable in the 4th mora, unaccented syllable in the 1st mora, or an accented syllable in the 2nd mora
3 = extrametrical syllable in the 5th mora, unaccented syllable in the 2nd mora, or an accented syllable in the 3rd mora
4 = extrametrical syllable in the 6th or later mora, unaccented syllable in the 3rd mora or accented syllable in the 4th mora

Translating cultures within the EU

Nicola Borrelli
University of Naples "Federico II"

Failure to ratify the European Constitution by popular vote in France and the Netherlands in June 2005 proved the strength of anti-globalising spurs within the EU. The attitude of each Member State towards the European Institutions depends on its national culture, and different languages shape the original message conveyed by the Union into different ideas.
The aim of this paper is to explore the way in which the language used in EU documents reflects the cultural expectations of the prospective audience in terms of both content and stylistic features. In particular, the analysis aims to identify any correspondence between the degree of Euro-scepticism among European citizens and the language used in a EU product.
In the European Union, translators play a crucial role in the communicative process between the EU institutions and citizens. Specific translators' choices can be the result of either their 'cultural frames', acting at the unconscious level, or their deliberate choice to shape the text in such a way as to suit their audience's specific communicative needs.
The analysis is carried out on the script of a video on the European Constitution released by the EU Audiovisual Service, and is grounded on concepts drawn from translation studies, as well as on Geert Hofstede's model of cultural dimensions.

1. Introduction

"United in Diversity": the motto of the European Union summarizes the paradox on which the world's largest confederation of independent states seems to be based. In fact, the effort to move towards political and economic harmonization made by the Union as a supranational institution appears to have to contend with the diverse cultural identities of each of its 27 member states. The use of different languages is the most tangible expression of the cultural heterogeneity of the European Union. With Romanian and Bulgarian getting official status in the EU – following the entry of the two Central Eastern European countries into the

Union on the 1st January 2007 – and the arrival of Ireland's national language on the same day, the number of the EU's official languages has risen to 23. Linguistic diversity has always been a key theme within the European Union: all official documents are translated into all official EU languages[1] and gain force of law in a member state when they are translated into the officially recognized language of that state (Williams 2002). It is crucial to understand that translation is not a word-by-word re-encoding of a text into another language (Taft 1981; Schäffner 1996; Katan 1999; Seleskovitch & Lederer 2001), but a culturally-mediated product aiming at raising in the target readership the same psycho-emotional and cognitive responses as in the readers of the source text (Salmon 2005). The translator is not simply a civil servant working for the European Union, but also a representative of the national culture of the country s/he is from. In his extensive study carried out across more than 50 countries, Hofstede (2001: 21–22) remarks that translators tend to filter meanings according to the dominant system of values of their countries, and the impact of culture begins in the translator's mind. At this point questions arise: does this culture-bound work of translation serve the politically and economically globalising purposes of the EU's official documents? May translators – consciously or unconsciously – let their systems of values and their cultural baggage interfere with their work and with the goals of the European Union? Such queries shed a disquieting light on the multilingualism of the European Union, and foster provocative challenges to the "diversity at all costs line" (Williams 2002: 45). This is the case with van Els (2001: 349) who asserts that

> It is a myth that the great diversity of languages and cultures as such is a good thing and that, consequently, its present manifestation in the EU represents a great richness, a treasure that should be defended at all costs. It is one of the myths that co-determine current EU policy on institutional language use.

This study aims to provide answers to the above questions. In particular, it seeks to find out to what extent the translations of Brussels' official documents mirror the specific national perspectives of their translators and how these localising spurs interact with the general policies of the European Union. It follows up on previous research work conducted in 2005 (Borrelli & Pistillo 2005), focusing in particular on those documents used by the EU for disseminating purposes. As a product carried out by a particular individual for his/her fellow citizens, translations are analysed in terms of cultural mediation, which can be *aware* or *unaware*, "[…] depending on whether certain translation choices can be considered as the

1. As for Irish, EU institutions are not required to translate all legislation into it, mainly for practical reasons.

result of the translators' cultural background, or of their deliberate choice to shape the text in such a way as to suit their audience's specific communicative needs." (Borrelli-Pistillo 2005: 15)[2]. Data supplied by the Eurobarometer show that different countries in the EU hold very different attitudes to the European institutions, resulting in an articulated spectrum that ranges from keen Euro-enthusiasm to open Euro-scepticism. What has increasingly drawn attention to these data is the most dramatic incident that has shaken the EU over the last few years: the negative return of the French and Dutch referenda for the ratification of the European Constitution.

2. Euro-enthusiasm vs. Euro-scepticism: The different spirits of the EU

The failure of the two popular referenda held in France and in the Netherlands in spring 2005 to ratify the European Constitution has presented the European Union with one of the most delicate moments in its history.

First and foremost, this incident has brought the ratification process to a halt, with many countries shelving the planned general vote to incorporate the newborn common Constitution into the national law. At present, referenda to ratify the proposed Constitution appear to be indefinitely postponed in the Czech Republic, Denmark, Ireland, Portugal and the United Kingdom[3].

Secondly, the rejection of the Constitution in two founding member states has revealed a feeling of Euro-scepticism lurking beneath the European Union.

Below is the graph from the survey Eurobarometer carried out in autumn 2006 that polled the citizens of the 25 member states of the EU – plus those of Romania, Bulgaria Croatia and Turkey – about their opinion on their countries' present, future or potential membership in the EU.

---

2. Here I am not arguing that aware and unaware translation choices can be told apart. I just mean to say that a translator might deem one word more appropriate than another when addressing a specific audience, or might unconsciously make a choice that betrays his/her specific national schemata. Hofstede (2001: 9–10) has defined values as "the core element of culture" and de Mooij (2005: 79) states that "values are among the first things children learn, not consciously but implicitly. Developmental psychologists believe that by the age of 10, most children have their basic value systems firmly in place". Thus, despite not being a *national product* in the narrowest sense of the term, an EU translator might still be affected by his/her cultural values.

3. In the United Kingdom the ratification procedure also included the parliamentary approval, but also this process was suspended by the UK government on the 6th June 2005. Further information and constant updates concerning the state of play of the common constitution ratification course are available, in the EU's 23 official languages, at the link www.europa.eu

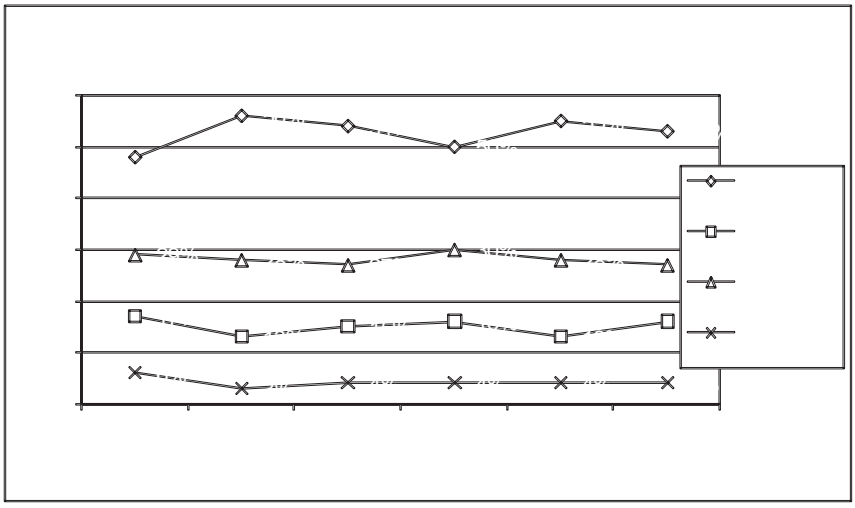

Figure 1. Adapted from Eurobarometer 66, Autumn 2006, Public Opinion in the European Union, 9

Figure 2. Adapted from Eurobarometer 66, Autumn 2006, Public Opinion in the European Union, 34

As shown in Figure 1, the number of respondents thinking of membership as a good thing slumped in the autumn 2005 after the referenda (50%; −4 points), and this occurred concomitantly with a rise in the number of interviewees deeming membership a bad thing (16%; +1) and with a surge of the number of those showing indifference towards membership (30%; +3). The most recent figures (autumn 2006) indicate a decrease in the number of the 'neither good nor bad' informants (27%; −1%), a similar reduction in the number of EU membership supporters (53%; −2%) and a rapid increase in the group of its opponents (16%; +3%). Apparently, those who made up or changed their minds between spring and autumn 2006 did so for the worse.

The broken down data provides interesting national results: in the Europe of the 25, Britons are the least supportive of the EU[4], with only 34 respondents out of 100 regarding membership as a good thing against a EU25 mean of 53%[5].

As far as the EU Constitution is concerned, the same Eurobarometer survey measured the degree of support for the constitutional text among citizens of both the countries that have not yet adopted it and those that have already rejected it.

Figure 2 shows that the public opinion appears more positive in autumn 2006, the European Constitution being supported by 53% of the respondents, with an improvement of 6 points compared with spring 2006. Nevertheless, the broken down data registers once again a very different situation in the UK.

Among the EU10 (Table 2), the member states where the ratification process is either pending or has failed, the United Kingdom is the one with the least respondents in favour of the European constitutional text (40%) as of autumn 2006, with a loss of two more points as compared to the figure reported in the EU25 survey from spring 2006 (Table 1)[6]. In addition, the gap existing between those in favour and those against the Constitution in the country is tiny (40% vs. 35%)[7].

Table 1 also shows that in spring 2006 the greatest support for the constitution was registered in Hungary, Belgium, Germany and Italy, four member states where a parliamentary ratification procedure has been adopted and no referendum has

---

4. The British percentage is also lower than those of Romania (62%) and Bulgaria (55%), which joined the EU on the 1st January 2007, and Turkey (54%), at present just a candidate country.

5. *Eurobarometer 66*, Autumn 2006, Public Opinion in the European Union, 10.

6. The spring 2006 survey (*Eurobarometer 65*) on the degree of support for the European Constitution is the latest one including data from all the 25 member states. The autumn 2006 survey (*Eurobarometer 66*) only measured the degree of support for the European Constitution in the 10 countries where the ratification is pending or failed (EU10). This accounts for the heterogeneity of the tables compared.

7. *Eurobarometer 66*, Autumn 2006, Public Opinion in the European Union, 35

Table 1. Adapted from Eurobarometer 65, Spring 2006, Public Opinion in the European Union, 24

Support for the European Constitution Spring 2006 Country results	
Hungary	78%
Belgium	75%
Germany	71%
Italy	71%
Slovenia	71%
Luxembourg	64%
Cyprus	64%
Spain	63%
Greece	62%
France	62%
Poland	62%
European Union (25)	61%
The Netherlands	59%
Lithuania	58%
Ireland	56%
Slovakia	55%
Portugal	53%
Czech Republic	52%
Latvia	52%
Estonia	47%
Malta	46%
Denmark	45%
Finland	45%
Austria	44%
United Kingdom	42%
Sweden	39%
* CY (tcc(= 59%)	
Other Countries	
Romania	63%
Croatia	60%
Bulgaria	59%
Turkey	45%

been held. Paradoxically, the European Constitution appears to be best perceived by ordinary citizens in those countries where there was no direct involvement of the population in its ratification.

The data provided above has been used in this paper for the linguistic and cultural analysis contained in the following sections.

Table 2. Adapted from Eurobarometer 66, Autumn 2006, Public Opinion in the European Union, 35

Support for the European Constitution Autumn 2006 Country results	
Poland	63%
Portugal	60%
The Netherlands	59%
France	56%
Ireland	56%
Finland	56%
European Union (10)	53%
Denmark	51%
Sweden	50%
Czech Republic	50%
United Kingdom	40%
Other Countries	
Romania	70%
Bulgaria	60%
Croatia	57%
Turkey	47%

3. The corpus and the theoretical framework

The attitude of each member state towards the European Union is shaped by culture. At the same time, though, the messages conveyed by the European institutions to the different member countries are also culture-bound, insofar as they are the product of the mind of a translator. In fact, the translation process implies an interpretation of the source text by the translator (Seleskovitch & Lederer 2001: 93) through a three-stage process: (1) the original discourse, (2) the deverbalization of meaning units and (3) the reformulation of these units through a new discourse in the target language. During deverbalization, the actual translation takes place in the translator's mind, and is determined by the fusion into a whole of the significations of the words and the cognitive complements. Lederer in Seleskovitch & Lederer (2001) conceives of two types of cognitive complement: one that accounts for the comprehension of what is explicit and implicit in the text and another depending on cognitive baggage and which is classified as encyclopaedic. This encyclopaedic knowledge contributes to determine each person's cultural frames, that "[…] perceptual window through which an individual defines him or herself, others, and the world." (Brake *et al.* 1995, quoted in Sandrelli 2005: 80) Cultural frames play a major role in the cognitive processes underlying any communicative event;

therefore their influence cannot be neglected in translation activities. An analysis of the impact of national cultures on EU's official translations implicitly brings to the fore another issue: that of the persistence and the strength of culture-driven, linguistically-determined *localising* components within an environment characterized by a constant language contact and potential globalising spurs, and that of their interaction – or interference – with the policies of the Union.

The text chosen for analysis is the script of a video released by the European Commission's Audiovisual Service in January 2005, to present and to promote the contents of the European Constitution with the press and the general public (see also Appendix). The 16-minute footage, produced in the 20 official languages of the EU starting from a French source text, was viewable at the section *mediatheque* of the Union's official website at least as late as September 2005[8]. A new search for the video in late 2006 yielded no results; therefore further inquiries into the matter were made with Jean-Pierre Assalone, the head of the e-team in charge of the EU's Audiovisual Service, as late as January 2007. The video in question was reported to have been withdrawn by the hierarchy of the Audiovisual Service for editorial reasons, and its distribution to have been stopped. Mr. Assalone also added that although an update of the video was foreseen, no date had been set yet.

The Italian and English translations of the original French script were picked for analysis in the present paper. The data provided in Section 2 shows that Italy and the United Kingdom stand at the two extremes of the Euro-enthusiasm spectrum, especially as far as support for the European Constitution is concerned. This consideration spurred a contrastive analysis between the English and Italian target texts[9] compared to the French original, with the aim of finding out whether the Eurobarometer data was mirrored in the translation choices made. The corpus was investigated within a theoretical framework including *translation theory* and Hofstede's (2001) model of *cultural dimensions*, which are discussed in detail in the following subsections.

3.1 Translation theory

The principle of *equation* is a fundamental one in translation theory. Taylor (1998: 49) defines it as "[…] the default position whereby if no other pressing

8. The video was last consulted on the 5th September 2005 at the link http://www.europa.eu.int/comm/mediatheque/video/constitution_fr.html

9. The Italian and English target texts were also chosen in consideration of the author's L1 and FL skills.

reason exists a term should be translated by its one-to-one equivalent". This implies that equation is one of the main strategies applied by translators in their work. Malone (1988) provides a concise and comprehensive description of translation methods, which is mainly based on the classic *Stylistique Comparée du Français et de l'Anglais. Méthode de la Traduction*, by Vinay and Darbelnet (1958; 1995). He argues that equation occurs when there is some sort of automatic equivalence, and he offers a convenient terminology for the description of the other techniques used in translation activities. *Substitution* is applied when there is no immediate correspondence between the source and target texts with respect to a specific item. In the *divergence* strategy, the translator chooses from a range of possible alternatives in the target language, whereas in *convergence* one expression in the target language can translate several of the source language. In *amplification* and *reduction*, the translator respectively adds to or eliminates from the original text elements s/he considers necessary or redundant. *Diffusion* and *condensation* concern the greater elaboration or tightening of source text expressions when translated into the target language. Finally, *reordering* is about acting at the level of syntax to create familiar patterns in the target language.

In the following analysis particular attention will be paid to those items of translation where equation was discarded in favour of a different technique without apparent restraints existing to bar the translator from using it. Having considered the likelihood of translation mistakes insignificant, given the generally great proficiency of the professional translators working for the EU institutions[10], these particular translation choices have been interpreted as a possible result of an *interference* of the translator's "linguoculture" (Salmon 2005), especially in terms of acceptance of the EU's policies. In other terms, the issue has been framed as a mainly cultural one, closely related to the system of values, behaviours, and beliefs of the translator and of his/her potential audience.

3.2 Hofstede's cultural dimension

In his highly influential *Culture's consequences: comparing values, behaviors, institutions, and organisations across nations*, Geert Hofstede (2001) identifies five independent dimensions of national culture differences, each of which is rooted in a basic problem with which all societies have to cope, but on which their answers vary. The dimensions are as follows:

10. In this respect, it is worth pointing out that official EU translators exclusively translate into their mother language.

1. *Power distance (PDI)*, which is related to the different solutions to the basic problem of human inequality;
2. *Uncertainty avoidance (UAI)*, which is related to the level of stress in a society in the face of an unknown future;
3. *Individualism versus collectivism (IDV)*, which is related to the integration of individuals into primary groups;
4. *Masculinity versus femininity (MAS),* which is related to the division of emotional roles between men and women;
5. *Long-term versus short-term orientation (LTO)*, which is related to the choice of focus for people's efforts: the future or the present.

Hofstede measures the degree to which each dimension is present in the different cultures by means of an index. For the purposes of the present study, the Power Distance Index (hereafter PDI) and the Uncertainty Avoidance Index (hereafter UAI) will be dealt with in greater detail, for two main reasons: firstly, because the cultures analysed appear to have very different scores on these two indices; secondly, because these two dimensions and their implications seem to be the most relevant to the type of issues found in the corpus.

Cultures scoring high on power distance are characterized by an uneven distribution of power. Society has a marked hierarchical structure, with the few people in power holding privileges and being treated with respect and obsequiousness by their subordinates by virtue of their supposedly greater wisdom and goodness. In cultures scoring low on power distance, by contrast, power is shared by many and is usually not paraded. In these cultures, the relationship between citizens and the institutions is one between peers. Citizens are not in awe of institutional authority and tend to take active part in the political and civil life of their communities.

The dimension of uncertainty avoidance is concerned with the way people deal with the fear of ambiguity in society and uncertainty about the future. All human societies have developed ways of coping with uncertainty, namely technology, law, and religion, but their attitudes towards uncertainty and the anxiety it causes vary.

Cultures with a high level of uncertainty avoidance view what is different or new as potentially dangerous; therefore they resist change and rely strongly on rules. In these cultures, citizens tend to delegate much of the decision-making power to the authority as a way to lighten the burden of responsibilities and thus lower anxiety levels. As a whole, cultures scoring low on the UAI appear to be much more tolerant of change and diversity. People tend not to be afraid of taking on their own responsibilities and wish to be an active part of the decision-making processes of their communities, without fearing the onset of higher levels of uncertainty and anxiety.

Hofstede's data (2001) show that Italy scores high on both the PDI and the UAI, indicating that it is a rather hierarchical society with a low level of tolerance towards uncertainty. The UK, in contrast, appears to have a much more equal citizen-institution relationship, as well as greater tolerance of change and uncertainty.

4. The main text and the *Vox Pop*

This section of the paper is devoted to the analysis of the corpus, as defined in paragraph 3 above. The script of the video considered consists of two parts: (1) the main text and (2) the *Vox Pop*. The subsections 4.1 and 4.2 below explore them in detail.

4.1 The main text: The linguistic analysis

The main text is a transcription of the narration made in the video by a voice over. For the purposes of this study, the Italian and English versions have been compared with each other and with the original French source. A qualitative analysis method has been adopted, selecting chunks of source text for which translation techniques other than equation had been made in either or both target language(s). Spatial constraints have not made it possible to consider all the instances of oblique translation present in the target texts, so the most interesting and representative ones have been chosen for analysis. The parts of the reported texts which have been made the object of analysis have been underlined.

Case 1

FR: Bon nombre d'européens s'interrogent en effet sur la portée de ce qui est appelé à devenir leur Constitution commune.

EN: Many Europeans are concerned about the real significance and consequences of what is destined to become their common constitution.

IT: Molti europei si interrogano infatti sul significato reale e sulle conseguenze di quella che è destinata ad essere la loro costituzione comune.

The difference between the Italian equation *s'interrogano* and the English substitution *are concerned* is interesting. It seems to indicate that, for the British translator, EU citizens – but they may well be the British citizens – are *worried* about the consequences of the introduction of the European Constitution, almost as if this might endanger an existing *status quo*. An equation would have been possible here: the British translator could have chosen the verb *to wonder*, but s/he opted for a verb connoting trouble or anxiety. This clashes with the French *s'interrogent* and the Italian *s'interrogano*, two neutral verbs that simply denote a questioning

process eliciting an answer, which may be either positive or negative. Looking at this lexical choice through the lens of Hofstede's cultural dimensions, it may appear a way to connote the uneasiness experienced by the low-PDI British culture in the face of the centralising potential of a European Constitution. Let us remember that the UK, as a Common Law country, is the only EU member that does not have a written constitution[11]. Furthermore, as a low PDI country, the UK appears to be characterized by an active cooperation between citizens and authorities, in the framework of a lesser centralization of political power and of an enhanced system of representation.

Case 1 offers another interesting example from this point of view. The Italian translation *significato* – intensified by the adjective *reale* – diverges from the source text *portée* despite the existence of the equivalent *portata*. The translator here chooses from a range of possible alternatives in the target language: he discards the options *portata* or *importanza*, and uses the term *significato*, which primarily means *meaning* and, only in a figurative way, denotes importance. In so doing, the Italian version seems to be more focused on the formal comprehension of the Constitution than on its possible implications. To focus on the aftermath of the introduction of a European Constitution would mean an increase in uncertainty levels, which is not desirable in a culture that ranks high on Hofstede's UAI, as is the case with Italy. Let us consider that Italy is among the countries which have chosen a procedure of parliamentary ratification of the European Constitution, with citizens not being directly involved in the process (see Section 2).

Case 2

FR: Aujourd'hui, la "machine" fonctionne. Mais elle est de plus en plus complexe. Un véritable dédale! Difficile d'en appréhender tous les méandres. Le mode d'emploi européen est de plus en plus lourd.

EN: Today, the "machine" is up and running. But it is increasingly complex – a real maze. It's difficult to understand all the ins and outs and Europe's «user manual» is becoming bigger and bigger.

IT: Oggi la "macchina" funziona, ma è sempre più complessa, un vero dedalo. È difficile comprenderne tutti i meandri. La guida all'uso dell'Europa è sempre più corposa.

Case 2 introduces us to the use of metaphors within the EU. All three languages bring in the metaphor of the *machine* to refer to the European Union, and resort to the additional one of the *maze* to symbolize its complexity. Nevertheless, the choice of the successive noun phrase *ins and outs* in the English text marks a significantly different translation choice with respect to the French and the Ital-

11. See also Borrelli & Pistillo (2005: 44).

ian. The *Online Oxford English Dictionary* defines the *ins and outs* as "windings or turnings in and out, devious or tortuous turns to and fro in a road, a course of action, etc.; sinuous ramifications". A similar definition is that given for the term *meander*: "a winding course", and, especially in plural, "a crooked or winding path (of a maze); a labyrinthine passage". Though, the idiomatic *ins and outs* has an additional meaning, , that of "complete details of a matter" (*The New Chambers Dictionary* 2003), which has a substantially positive connotation. Choosing this phrase, typical of a user-manual-like, informal, register, the British translator foregrounds the initial mechanical metaphor. The Union is presented as a complex machine, but a user manual exists to comprehend it, and the translator seems to be inviting his/her low-PDI fellow citizens to consult it before deciding whether to, so to say, buy it or not. In other words, the EU and its complex procedures are part of the uncertainty of life with which British society optimistically copes. This is a rather different perspective from that of France and Italy, two high-PDI countries with a very low tolerance of uncertainty. Here the Union – and metonymically all the treaties on which it is based – are mainly compared to a *dédale / dedalo* full of difficult-to-comprehend *meanders / meandri*, and are therefore better left for the institutions to deal with. The focus is on the metaphor of the maze more than on that of the machine; a user manual also exists here, but it is a *guida corposa / mode d'emploi lourd*, an unwanted burden.

Case 3 lingers on the use of metaphors:

Case 3
FR: Aujourd'hui, l'Union est régie par 8 traités.
EN: Today, the Union is governed by 8 treaties.
IT: Oggi l'Unione si fonda su otto trattati.

The French verb phrase *est regié* is translated into English with the equation *is governed*. The verb *régir* is defined as "to govern, to lead" (*Le Petit Robert 2004*), and the main entry for *to govern* is "to officially and legally control a country" (*The Longman Dictionary of Contemporary English* 2003). In both cases, the semantic scope of the machine metaphor introduced in Case 2 is broadened to encompass the idea of a bureaucratic machine, made of the treaties which keep the EU together. Differently from the English, the Italian discards the equation (which would have been perfectly possible through the use of either the verb phrase *si regge* or *è retta*), opting for the substitution *si fonda*. Now, the primary meaning of the verb *fondare* is "to start building by laying the foundations" and, in a wider sense, "to base, to centre" (*Dizionario De Mauro Paravia Online*). This means that the reflexive form *fondarsi* denotes the idea of the Union *being based on* eight treaties, but, at the same time, connotes the Union as a *building* endowed with foundations. In other terms, the Italian choice marks a departure from the

metaphor of the machine, to call in the more reassuring metaphor of the *house*, one particularly welcome in a high-UAI society.[12]

The same translation strategy can sometimes serve different purposes, as proved by Case 4:

Case 4
FR: A 25, le vote à l'unanimité devient difficile.
EN: With 25 members, unanimous voting becomes difficult.
IT: Con 25 paesi, il voto all'unanimità diventa difficile.

Here both English and Italian resort to amplification to make the noun phrase *A 25* of the source text more explicit. Nevertheless, whereas the British translator opts for *members,* his/her Italian colleague uses the term *paesi*. A quantitative review on the texts is given in Table 3:

Table 3. Quantitative analysis of the usage of amplification of A 25

	membre(s)/état(s) member(s) member(s)/member state(s) membro(-i)/stato(-i) membro(-i)	**Pays Country(-ies) Paese(-i)**	**Total**
French	13	3	16
English	18	3	21
Italian	13	7	20

In the French source text, it is shown that the noun phrase *membre(s)/état(s) member(s)* is used thirteen times[13], whereas *pays* is used three times, and all the instances of both terms are equated both in English (*member(s)/member state(s)* and *country(-ies)*) and in Italian (*membro(-i)/stato(-i) membro(-i)* and *paese(-i)*). Nonetheless, the total number of occurrences of *member(s)/member state(s)* in

12. The IDV dimension could also be significantly involved in the choice of the house metaphor in the Italian translation. Hofstede (2001) has listed Italy among the high-IDV countries, but his data derived exclusively from informants from northern Italy. Later studies, based on north-south aggregate data (and agreed on by Hofstede), have shown that Italy is actually a country verging on collectivism (De Mooij 2005). Among the key characteristics of collectivistic cultures, Hofstede (2001: 236) has mentioned the importance of the family as an institution providing "[…] protection in exchange for lifelong loyalty". Now, if we think that the idea of a house is closely connected with that of a family, the shift to a homely metaphor in the Italian translation appears quite culture-bound.

13. This figure does not include the cases in which the noun phrase *membre* has been used to refer to something else from the EU member states (e.g. "membres du Conseil").

English is eighteen[14], and the total number of occurrences of *paese(-i)* in Italian is seven[15]. This implies that five more times in English and four more times in Italian, the translators refer explicitly to the EU states while the French source text does not. Though, whereas the English resorts to the noun *member(s)/member state(s)*, the Italian prefers *paese(-i)*. The amplifications in Case 4 are an example in this sense. The technical and more context-specific *members* conveys a connotation of the EU as a political, artificial entity, showing the English translator's greater sensitivity to the scope of the Union and the possible consequences of its policies in a low-UAI country where there is less support for European government. Conversely, the more generic *paesi*, normally used to denote a geographical space, gives an idea of the Union as a community of countries: there is no mention of the EU as an economic-political ensemble, perhaps because of the weaker interest of high-PDI Italians in politics, or arguably because guidance from above is welcome more than feared in a high-UAI country like Italy.

However, the situation appears to be inverted in Case 5:

Case 5

FR: L'Europe écoute-t-elle les préoccupations de ses citoyens? Comment ceux-ci peuvent-ils encore se faire entendre des Institutions dans une Europe à 25 ?

EN: Does Europe listen to its people? How can individuals make themselves heard by the institutions in a Europe of 25?

IT: L'Europa ascolta i suoi cittadini? Come si può far sentire la propria voce alle istituzioni di un'Europa a 25?

Whereas the Italian version prefers *cittadini* as an equivalent of the French *citoyens*, the English version opts for *people*. A *people* is a "body of persons held together" firstly "by a common origin, speech, culture", and only secondly by "political union, or […] a common leadership" (*The Chambers Dictionary* 2003). This lexical choice – apparently contradicting what was argued in Case 4 – can be better explained by means of a linguistic and a political remark. A *citizen* is a "member of a state" where a *state* is defined as "a political community under one government" (*Chambers Dictionary 2003*). A citizen enjoys a number of rights insofar as s/he belongs to a country run by a specific government; in other terms, as a lexical item, *citizen* sounds much more politically connoted than *people*. Among

14. Obviously, the cases in which the noun phrase *member* has been used to refer to something else from the EU member states (e.g. "Council members") have also been excluded from the count of these occurrences.

15. Conversely, the occurrences of *membro(-i)/stato(-i) membro(-i)* in Italian, and the occurrences of *country(-ies)* in English are exactly as many as the occurrences for *membre(s)/état(s) member(s)* and *pays* in French (i.e. thirteen and three respectively).

the rights of a citizen is obviously the right to vote. In the UK only individuals whose name appears on the electoral register are entitled to vote.[16] We could say therefore that a British national can exercise full citizenship and cast his/her vote provided s/he has registered to do so. Evidently, the status of *citizen* presents many more constraints than that of *people*. Arguably, from a British perspective, a European Union speaking to *citizens* may not be reaching everybody, cutting out those individuals that, for various reasons, are part of the *people* though not being fully entitled to being called citizens. Obviously, this would clash with the effort to minimize inequality in society, which is a distinctive feature of relatively low PDI countries (Hofstede 2001). Conversely, we could speculate that the more *intimate* tone conveyed by the noun *people* would be regarded as less authoritative and credible by a high PDI-culture (Evans et al. 2004: 135–139).

We can notice that neither the English nor the Italian version provides a translation of the French *préoccupation*. In the wake of the cultural considerations made so far, here also a different perspective of the EU-member states relationship may underlie the same strategy of reduction performed by the two languages: whereas in English it may demand exchange on a more comprehensive range of issues on behalf on the low- UAI British society, in Italian it may just omit a term likely to cause alarm within the high UAI Italian culture about matters that are generally delegated to institutions.[17]

Case 6
FR: L'Europe seraitdonc au seul service du marché […].
EN: So will the European Unionsimply be a servant of market forces […]?
IT: Quindi l'Europa sarebbe solo al servizio delle forze di mercato […].

Interestingly, here the three languages resort to different verb moods to express the same concept. Whilst the Italian translation, following the French source text *serait,* uses a conditional (*sarebbe*), the English one prefers the modal *will*. Furthermore, whereas a direct question is formulated in the English translation, this is not the case with the Italian version and French source text. The English text seems to express real reservations about the nature of the EU, and to invite its low- PDI and low-UAI prospective audience to decide autonomously as to whether Europe

16. For further information on the British voting system you can visit the link http://www.direct.gov.uk/RightsAndResponsibilities/RightsAndResponsibilitiesArticles/fs/en?CONTENT_ID=10014442&chk=tW2II/

17. Hofstede (2001) writes that in low PDI and low UAI countries citizens cooperate with political authorities (p. 116) and show a strong interest in politics (p. 180). Conversely, high PDI and high UAI societies tend to wait for action by authorities (p. 116) and are characterised by a weak interest in politics (p. 180).

could be a *servant of the market forces*. However, the Italian and French scripts opt for the conditional mood, which gives a rhetorical tone to the question, thus making their higher-PDI and higher-UAI audiences expect a negative answer.

4.2 The VOX POP: The linguistic analysis

In the video investigated in the present analysis the flow of the main text is interspaced by interviews with ordinary citizens from diverse countries in the EU. They form the section called *Vox Pop*.[18] The respondents answer questions concerning some key issues of today's EU in their native language, and subtitles in the main tongue of the video are provided[19].

Bearing in mind the technical constraints that are applied to subtitles[20], the same theoretical framework used for the analysis of the main text has been adopted here. The author has focused on those choices in the source text – obviously not always coinciding with French here – and in the Italian, English and French versions that could not be exclusively ascribed to translation praxis. The different translations have been compared to each other and to the original message; the latter has been marked in each case by the acronym SL (Source Language) in brackets.

Case 1/VP

FR: Ilfaut une constitution, sans quoi l'Europe serait comme une maison sans toit.

EN: Thereshould be a constitution otherwise the whole thing becomes a headless system.

IT: Ci vuole una costituzione altrimenti l'Europa sarebbe una casa senza tetto.

DE: Ja, also eine Verfassung sollte natürlich schon sein, weil sonst ist die ganze Sache ein, ein System ohne Kopf, ehrlich [...] (SL)

This case displays a twofold contrast between the "linguocultures" (Salmon 2005) of two countries with a high PDI (France and Italy) and two with a low PDI (the

18. *Vox Pop* is the name of the interview section in both the English and the Italian translations. In the French source text it is named Micro Trottoir. Cases in this subsection will be marked by the letters VP (Vox Pop) to distinguish them from the main text analysed in subsection 4.1.

19. Subtitles are absent when the interviewee's language coincides with that of the main text.

20. Gottlieb (1993: 164) cites a Swedish finding dating back to early seventies, that average television viewers needed 5–6 seconds to read a two-liner of some 60–70 characters. He goes on to mention a Belgian study, according to which subtitle reading could be faster. Generally speaking, these findings have to be taken with caution, because subtitle reading speed can be influenced by a number of variables (e.g. the type of medium for which the subtitles are produced, their semantic and syntactic complexity and, last but not least, the fact that subtitle reading is not the prime interest of viewers).

UK and Germany). The English translation is the exact equivalent of the original German text: the modal *should* translates the German *sollte*; the *whole thing* and the *headless system* correspond to *die ganze Sache* and *ein System ohne Kopf* respectively. Both the German interviewee and the English translator define Europe as a *thing*, which – without a Constitution – might even turn into a *headless system*. It is also noticeable that the metaphor used for Europe is that of a *system*, "a set of connecting things or parts forming a complex whole" (*The New Oxford Dictionary of English* 1998), which admittedly recalls complicated bureaucratic machinery. Nevertheless, a Constitution appears to be no more than advisable in the two languages, as proven by the fact that English also equates the German source text in the use of the verbal mood (*there should be / sollte … sein*).

French and Italian have recourse to a completely different system of connotation, by referring to the European Union simply as *l'Europe / l'Europa*. By opting for the generic term instead of the more technical *Union* these two languages merge the two concepts of Europe as a political and cultural whole. The need for a Constitution is clearly expressed through the verbs *il faut / ci vuole*. It may be worth considering this verb pair, as opposed to the previous one (*there should be / sollte…sein*) both in terms of semantics and of modality. Firstly, French and Italian choose verbs that intrinsically express a need – *falloir* and *volerci* respectively – whereas the original German and the English translation use *sein* and *to be*. Secondly, the choice of the indicative mood in the two Romance languages marks the clauses as *Realis* in contrast to the Germanic pair which can be categorized as *Irrealis* both for the use of modals and for the use the conditional mood, according to the definition of Palmer (2001: 4) whereby the distinction between the two categories "[…] depends on the distinction between what is asserted and what is not". It is also interesting to notice the shift in the metaphorical representation of the EU: the disquieting metaphor of the *headless system* is replaced in French and Italian by the more reassuring one of the *maison sans toit / casa senza tetto*. Indeed, the welcoming image of the house is used in a variety of contexts in EU texts and legislation and has proved to be a very rich metaphor, as it has inspired many other images related to the European space (Caliendo forthcoming).

Case 2/VP

FR: Les procédures de décision, c'est la bouteille à encre pour les gens ordinaires comme moi.

EN: Decision procedures, these are one black box for the ordinary people like me. (SL)

IT: Il processo decisionale è la bestia nera di noi persone comuni.

Interestingly, the three languages have recourse to different metaphors to refer to the EU decision procedures. In English – which is also the source language in this case – decision procedures are a "black box". A *black box* is defined as "Any complex piece of equipment *with contents which are mysterious to the user*" [emphasis added] (*The New Oxford Dictionary of English* 1998), used also figuratively to indicate anything complex and incomprehensible. A similar connotation is present also in the French solution *bouteille à encre*, which maintains the idea of mystery and lack of clarity, denoting "a complex or obscure matter" (*Le Petit Robert* 2004). In other words, the recurring metaphor of a complicated piece of machinery is introduced in the script once again.

The Italian resorts to a metaphor belonging to a completely distinct domain: *bestia nera*, that is, "a thing or a person one cannot stand, overcome or dominate" (*Dizionario d'Italiano Garzanti Online*), whose closest translation into English is a *pet hate*, i.e. "an intensely disliked person or thing" (*The Oxford Dictionary of English* 2003). Such a substitution implies no desire to comprehend EU's decision procedures: an object of dislike is something one tends to avoid, not to look into. Much better to leave it to be cared about by specialists. High PDI and high UAI appear to be at work again here.

Case 3/VP
FR: Tout est controlé. Ce sont des maniaques du contrôle.
EN: Everything is controlled, they are control freaks. (SL)
IT: Tutto controllato. Sono dei controllori nati.

The style and tone of the English text are deeply altered by the replacement of *freaks* with *controllori nati*. Not only is the Italian register more formal, but also its general overtone appears smoother. Two possible equations are discarded (*fissato* or *maniaco* would have served the purpose perfectly) in favour of the noun phrase *controllori nati,* literally *natural-born controllers*. Admittedly, the negative connotation implicit in *freaks* almost disappears in the Italian version, and this is not surprising for a culture where control *from above* is viewed as a welcome guidance rather than a limit.

Interestingly, both in case 2/VP and 3/VP the French translation – that is not the source text here – appears to be more consistent with the English original, whereas in case 1/VP it shared with Italian the same connotative system of the EU through the use of the same metaphor. It would seem that the French translator, though accepting the idea of the European Union as a *common house*, becomes more wary when the issues at stake are the degree of national participation in the communal decision procedures and the limitation of some national liberties through centralized control. This idea would seem to be suggested also by other

answers provided by French interviewees in the Vox Pop section, such as Nobody can really do what they want.[21]

The paternalistic image of the EU as a house whose hierarchy provides guidance and takes care of complicated issues is that conveyed by the Italian version of the Vox Pop, and arguably it proves to be the most culturalized one. The contributions of the two Italian respondents interviewed would seem to reinforce this hypothesis: I think more power should be given to the European Union and I think that a Minister of Foreign Affairs at European level for the whole Union would be a good idea. Apparently, both the language and the statistical data bear out the fact that Italy is among the member countries holding the most enthusiastic views on the EU, and this is particularly true of Italian public opinion. Nevertheless, Risse (2000) writes that Italy is one of the member states with the worst compliance records with law and regulations of the Union, whereas the UK is reported to have one of the best. How can this paradox be explained? Risse (2000: 3) reminds us that

> Groups of individuals perceive that they have something in common on the basis of which they form an 'imagined community' (Anderson 1991) […] Individuals frequently tend to view the group with which they identify in a more positive way than the "out-group".

From this perspective, Italian Euro-enthusiasm could be regarded as the result of a generic, historically inspired identification with the *imagined community* of Europe more than with the Union as a social-political organization. This does not appear unlikely in a country of relatively recent unification like Italy, where the very concept of national identity is, under many aspects, more *imagined* and politically devised than based on a factual common heritage subsuming all the cultural and social realities of the different regions.

5. Conclusions

In this study I have focused on the impact of national cultural frameworks on official translations within the EU, with specific regard to a document aimed at disseminating knowledge and acceptance of the European Constitution. The analysis conducted has shown how translators, while striving for fidelity to the source text and its purpose, cannot escape the influence of their culture. This inevitably determines the interpretation of the text and its translation in a way which seems to mirror the public opinion of the country targeted, as measured by the regular surveys conducted by the Eurobarometer. It remains difficult to tell whether the

21. French as a target language is not the aim of this study. These remarks are therefore sparse and tentative, and certainly worth proper investigation from specialists of French language and culture.

translator's cultural mediation is more aware or unaware, that is whether s/he is subconsciously influenced by his/her cultural baggage or intentionally shapes the text in such a way as to meet the communicative needs of the prospective audience. As written above, the EU's translators translate into their native language for their fellow citizens, so the demarcation line between what is consciously done and what is not may be rather unclear also in their own minds.

The analysis has also shown a correlation between translation choices, statistical data concerning the degree of Euro-enthusiasm of EU's members states, and national cultural dimensions as measured by Geert Hofstede. The Italian Euro-optimism emerged by the Euro-polls has proved to be matched by linguistic choices in the Italian translation that seem to point to a greater tendency to delegate decision-making power to an authority acting from above in order to reduce anxiety level in society. The Italian support for the EU does not seem to go hand in hand with a will to thoroughly understand its mechanisms, seemingly the result of a more general identification of the EU with a geographical entity.

Similarly, the Euro-scepticism coming out from the Eurobarometer data for the UK has found correspondence in the English translation, characterized by language betraying doubt about the benefits of the European Constitution – and more generally about EU membership – and concern over the effects of EU policy on the national law and on the long tradition of independence of the British people. Nevertheless, the British scepticism appears of a constructive kind, in that it implies a will to better understand the way the EU works: this shows linguistically through the greater lexical and syntactic exactness of the English text as compared to the French and Italian ones, and confirms a political and civil commitment related to its low ranking on both the PDI and the UAI.

Finally, in the light of the above findings, this paper has also revealed the localising potential of linguistic-cultural choices in an otherwise trans-national environment like the EU. In the translation of the script of the video promoting the European Constitution, culture has proved to outweigh language contact and to stand in the way of the unitary goal inspiring the text. If we can argue that a synergy has been established between the institutional aim of the EU and the Euro-enthusiastic filter of the Italian translation, a friction between the Union's endorsing purpose and the English translation also seems to have occurred. Ironically, the cultural diversity celebrated by the EU's motto in the text of the European Constitution may have hindered its ratification process.

The release of an update of the video analysed could be a starting point for an interesting follow-up, with the main aim at discovering if, and if so how, the failure of the 2005 referenda in France and in the Netherlands has determined a re-adjustment between institutionally-globalising and culturally-diversifying spurs within the EU.

References

Bochner, S. (ed.). 1981. *The Mediating Person: Bridges between Cultures*. Cambridge MA: Schenkman.

Borrelli, N. & Pistillo, M. G. 2005. Europe and 'the 25': Translating Cultures in EU Official Documents. In *'Languaging' and Interculturality in EU Domains*, G. Di Martino & V. Polese (eds), 11–65. Napoli: Arte Tipografica Editrice.

Caliendo, G. Forthcoming. The perception of boundaries: Doors and walls of the European common house. In *Atti del XXII Convegno AIA Cityscapes – Islands of the Self*.

De Mooij, M. 2005. *Global Marketing and Advertising – Understanding Cultural Paradoxes*. Thousand Oaks CA: Sage.

Evans, M. et al. 2004. Has the tone of online English become globalized? An empirical research study investigating the written tone of university web sites around the world. In *Proceedings of Cultural Attitudes towards Technology and Communication*, C. Ess & F. Sudweeks (eds), 135–139. Perth: Murdoch University.

Gottlieb, H. 1993. Subtitling: People translating people. In *Teaching Translation and Interpretation* 2: *Insights, Aims, Visions*, C. Dollerup & A. Lindegaard (eds), 261–274. Amsterdam: John Benjamins.

Hofstede, G. 2001. *Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organisations across Nations.* Thousand Oaks CA: Sage.

Katan, D. 1999. *Translating Cultures: An Introduction for Translators, Interpreters, and Mediators*. Manchester: St. Jerome.

Malone, J. L. 1988. *The Science of Linguistics in the Art of Translation*. Albany NY: State University of New York Press.

Palmer, F. R. 2001. *Mood and Modality*, Cambridge: CUP.

Risse, T. 2000. Regionalism and collective identities: The European experience. Prepared for the Workshop *El Estado del Debate Contemporáneo en Relaciones Internacionales*. <http://www.atasp.de/downloads/BuenosAires.pdf> (31 January 2007).

Salmon, L. 2005. Su traduzione e pseudotraduzione, ovvero su Italiano e pseudoitaliano. In *L'italiano delle traduzioni*, A. Cardinaletti & G. Garzone (eds), 17–33. Milano: Franco Angeli.

Schäffner, C. 1996. Translation as cross-cultural communication. In *Language, Culture and Communication in Contemporary Europe*, C. Hoffmann, (ed.), 152–164. Clevedon: Multilingual Matters.

Seleskovitch, D. & Lederer, M. 2001. *Interpréter pour Traduire*. Paris: Didier Erudition.

Taft, R. 1981. The role and personality of the mediator. In *The Mediating Person: Bridges between Cultures*, S. Bochner, (ed.), 53–88. Cambridge: Schenkman.

Taylor, C. 1998. *Language to Language*. Cambridge: CUP.

Van Els, T. J. M. 2001. The European Union, its institutions and its languages: Some language political observations. *Current Issues in Language Planning* 2 (4): 311–360.

Vinay, J.-P. & Darbelnet, J. 1958. *Stylistique comparée du Français et de l'Anglais. Méthode de la traduction.* Paris: Didier. English trans. and ed. by J. C. Sager, & M. J. Hamel. 1995. *Comparative Stylistics of French and English: A Methodology for Translation*. Amsterdam: John Benjamins.

Williams, C. H. 2002. Language policy issues within the European Union: Applied geographic perspectives. *Geografija in Njene Aplikativne Možnosti, Dela* 18: 41–60.

Dictionaries

The Chambers Dictionary. 2003. Edinburgh: Chambers Harrap Publishers.
The Oxford Dicitonary of English. 2003. Oxford: OUP.
The New Oxford Dictionary of English. 1998. Oxford: OUP.
The Online Oxford English Dictionary. www.oed.com
The Longman Dictionary of Contemporary English. 2003. Harlow: Pearson Education.
Le Petit Robert. 2004. Paris: Le Robert.
De Mauro, Tullio. 2000. *Il dizionario della lingua italiana per il nuovo millennio*, Torino: Paravia.
De Mauro Paravia Online. Il dizionario della lingua italiana. www.demauroparavia.it
Dizionario d'Italiano Garzanti Online. www.garzantilinguistica.it

Websites

Geert Hofstede's website: www.geert-hofstede.com
The European Commission's Audiovisual Service:
http://www.europa.eu.int/comm/mediatheque/video/constitution_fr.html
Eurobarometer: http://europa.eu.int/comm/public_opinion/index_en.htm

Appendix

THE SCRIPTS

1. The French script

Pourquoi une Constitution européenne?

Rome, 29 octobre 2004.
Les 25 signent le Traité instituant la Constitution européenne.
Ce texte qui modernise l'Union et le fonctionnement de ses Institutions se trouve aujourd'hui au centre des débats.
Bon nombre d'Européens s'interrogent en effet sur la portée de ce qui est appelé à devenir leur Constitution commune. Mais qu'en pensent-ils et qu'en attendent-ils?

MICRO TROTTOIR 1 :	
Des grands principes, c'est quelque chose qui manque.	FR
Sortir de cette simple idée que l'Europe est un vaste marché.	FR
Ca serait une bonne chose si l'Europe apportait la paix et la croissance pour tous ses membres.	FR
Nous avons besoin d'une constitution européenne pour savoir ce qui nous attend et de quoi il retourne.	SK
Il faut une constitution, sans quoi l'Europe serait comme une maison sans toit.	DE
Je ne pense pas que la constitution européenne dévalorise la constitution slovaque.	SK

Au départ, il y avait le charbon et l'acier.
Puis, l'Union européenne s'est construite, par étapes.
Aujourd'hui, la « machine » fonctionne.
Mais elle est de plus en plus complexe.
Un véritable dédale !
Difficile d'en appréhender tous les méandres. Le mode d'emploi européen est de plus en plus lourd.
Aujourd'hui, l'Union est régie par 8 traités.
Ils se sont empilés au fil du temps: de Rome à Nice, en passant par Maastricht, l'Acte unique, Amsterdam...
La Constitution les remplace tous. Plus simple et plus claire, elle devient la **base unique** de l'Union.
Pour commencer, ce texte définit les **objectifs** de l'Union.
Vivre en paix dans un espace de liberté, de sécurité et de justice. Créer un environnement économique stimulant, promouvoir la recherche, le développement durable, assurer un bon niveau de vie pour tous, …

La Constitution conditionne ensuite l'appartenance à l'Union au respect de **valeurs.** L'égalité des chances, la tolérance, la solidarité, le respect des minorités …

Plus loin, elle consacre également la **citoyenneté européenne.**

Elle garantit des libertés supplémentaires, comme celle de circuler et de s'établir partout dans l'Union européenne. Tout ressortissant d'un Etat membre en bénéficie, en plus de sa propre nationalité.

La Constitution reprend enfin la **charte des droits fondamentaux de l'Union.**
Des droits civils, sociaux, politiques ou religieux, voire administratifs ou bioéthiques, dont bénéficie toute personne vivant sur le territoire de l'Union.

MICRO TROTTOIR 2 :	
Les procédures de décision, c'est la bouteille à encre pour les gens ordinaires comme moi.	EN
On ne sait pas vraiment qui fait quoi.	DE
Nous avons nos élus à Bruxelles. A eux de nous dire sur quoi et comment ils votent.	SK
L'Europe a tellement grandi que les règles ne sont plus adaptées.	EN
Je me demande si les rapports de force sont équilibrés.	DE
J'ai l'impression que certains Etats ont plus de poids que d'autres. Grâce à la constitution, nous pourrons mieux décider ensemble.	SK
Si on met le même poids politique au 25, on n'y arrivera jamais. C'est pour cela qu'il faut dépasser un petit peu l'idée de la nation, pour penser vraiment "Europe".	FR

La formation européenne s'est considérablement agrandie.
Accorder un tel ensemble demande des règles claires et précises.
Sans quoi, on risque les couacs et le blocage de toute décision.
La Constitution clarifie donc le rôle et la composition des Institutions européennes et elle en modifie certaines règles de fonctionnement.

Le Conseil européen tout d'abord.
Avec la Constitution, il devient une Institution à part entière.
Tous les trois mois, les chefs d'Etat et de gouvernement s'y accordent sur les grandes orientations politiques de l'Union.
En fonction de ces orientations, c'est alors à la **Commission** de prendre des initiatives concrètes.
C'est elle qui prépare notamment les projets de lois européennes en toute indépendance et en veillant aux intérêts de l'Union.
Il y a aujourd'hui un Commissaire par Etat membre.
Pour garantir le bon fonctionnement de la Commission, dès 2014, date à laquelle de nouveaux pays auront normalement rejoint l'Union, la Constitution limitera le nombre de Commissaires à 2/3 du nombre des Etats membres, avec un système de rotation.
Le Conseil des Ministres représente, lui, les intérêts des Etats membres. On y adopte ou non les projets proposés par la Commission.
En fonction de la matière traitée, chaque pays y délègue son ministre responsable.
A 25, le vote à l'unanimité devient difficile.
A part dans certains domaines sensibles – la fiscalité par exemple – la Constitution généralise donc le recours à la majorité qualifiée.
Pour atteindre celle-ci, il faudra recueillir les voix d'au moins 55% des membres du Conseil représentant 65% de la population. Cette formule témoigne de la double légitimité de l'Union, fondée à la fois sur les Etats et sur les peuples.
Elu au suffrage universel tous les 5 ans, **le Parlement européen** représente les intérêts des citoyens.
La Constitution limite le nombre des députés à 750. Avec un minimum de 6 députés pour les petits pays et un maximum de 96 pour les grands.
Mais la Constitution consacre surtout le rôle de législateur du Parlement, aux côtés du Conseil des Ministres.
Quasi plus aucun projet présenté par la Commission ne pourra être adopté par le Conseil sans l'aval du Parlement.
Quant à la **Cour de Justice**, elle contrôle le respect de la législation européenne par tous, y compris par les institutions de l'Union.
Le droit européen comporte un éventail de plusieurs dizaines de sortes d'actes législatifs différents. Dans un souci de clarté, la Constitution les réduit à 6 : de la loi européenne, à la simple recommandation.

MICRO TROTTOIR 3 :	
Je trouve qu'il faudrait accorder encore plus de pouvoir à l'Union européenne.	IT
L'Europe des peuples est à construire, l'Europe sociale aussi et c'est ce qui importe.	FR
Le chômage. La justice. La fiscalité. Elle devrait être harmonisée dans toute l'Europe.	SV
Plus personne ne peut faire ce qu'il veut vraiment chez lui.	FR
C'est ça le paradoxe, c'est qu'il faut faire l'unité mais que chacun garde bien ces coutumes.	FR
Loi après loi, règlement sur règlement.	EN
Tout est contrôlé. Ce sont des maniaques du contrôle.	EN

L'Europe serait donc au seul service du marché…
Un prédateur, avide de compétences…
Une menace pour la souveraineté de ses membres…
Trois principes l'interdisent dans la Constitution.
Tout d'abord : l'Union ne peut agir que dans les domaines que les Etats membres lui ont expressément confiés.
Ensuite : l'Union ne peut intervenir que si son action apporte une valeur ajoutée par rapport à celle de ses Etats membres pris individuellement.
Enfin : l'Union ne peut dépasser ce qui est strictement nécessaire pour atteindre les objectifs fixés.
Mais, précisément, quelles sont les compétences de l'Union ?
UN : des « **compétences exclusives** ». La politique monétaire relative à l'euro ou les règles relatives à la concurrence dans le contexte du marché intérieur par exemple. Ce sont des matières où seule l'Union européenne peut agir.
DEUX : des « **compétences partagées** » telles que : l'agriculture, la pêche, l'énergie, les transports, la protection de l'environnement.
Autant de domaines dans lesquels l'Union ET les Etats membres peuvent prendre des initiatives complémentaires.
TROIS : Dans les matières telles que la santé, l'éducation, la culture, la politique industrielle, … les actions de proximité restent préférables mais l'Union peut néanmoins mener des actions **d'appui** ou de **coordination**.

MICRO TROTTOIR 4 :	
Il faudrait vraiment une politique extérieure européenne, ça c'est évident.	FR
Il faudrait une politique commune en matière de défense.	EN
L'intérêt national devrait s'effacer car le monde est un grand village.	DE
En ce qui concerne le Kosovo, l'Europe agit d'une seule voix. C'est là qu'ils peuvent montrer que ça sert à quelque chose.	DE
Je trouve que ce serait une bonne chose d'avoir un ministre des Affaires étrangères pour l'ensemble de l'Union.	IT
Je ne sais pas si on l'appellerait un président ou pas, mais une personne référente, qu'on sache quand on parle d'Europe, c'est telle personne.	FR

Les événements l'ont démontré, l'Union européenne doit accroître son influence et acquérir plus de poids sur la scène internationale.
« Europe ? Sure, what phone number ? » ironisait un jour Henry Kissinger.
La Constitution lui apporte une réponse claire.
Elle crée en effet le poste de **Ministre européen des Affaires étrangères**:
Nommé pour 5 ans par le Conseil européen, il sera la voix et le visage de l'Union vis-à-vis de l'extérieur.
Il conduira la politique étrangère et de sécurité commune de l'Union.
Les Etats membres conserveront leurs prérogatives dans ces domaines, mais le Ministre suscitera les prises de position communes.
Autre innovation qui renforcera le rôle de l'Europe dans le monde : la Constitution dote enfin l'Union d'une personnalité juridique.

Celle-ci est indispensable notamment pour siéger dans les instances internationales au nom de l'Europe.
On pourrait donc voir un jour l'Union siéger au Conseil de Sécurité des Nations Unies au nom de ses membres.

MICRO TROTTOIR 5 :	
Je pense que mon fils aura plus de possibilités que moi. *Je suis tout à fait optimiste sur ce plan.*	SK
Ils devraient commencer par simplifier le système d'échange d'étudiants.	SV
On ne retrouve plus tellement ces valeurs, on perd un peu de sa personnalité aussi.	FR
Ils font des lois stupides. *Ils nous enlèvent notre indépendance.*	EN
J'aime l'idée de la monnaie unique.	EN
– Et le progrès des droits de l'homme.	EN
C'est les fromagers qui ne vendent plus leurs fromages, c'est les usines qu'on fait fermer parce qu'elles ne sont plus à des normes européennes, c'est un scandale.	FR
S'ils veulent partir, s'ils veulent vraiment partir, allez-y, ils peuvent y aller…	FR
C'est une dictature. Je ne veux pas vivre dans une dictature.	EN
On ne peut pas avoir un pied dedans et l'autre dehors.	EN

L'Union est ouverte à tous les Etats européens qui partagent ses valeurs et ses objectifs.
Mais elle n'entend pas devenir un carcan.
Aussi, la Constitution prévoit dorénavant la possibilité de se retirer de l'Union.
Cette clause rappelle que l'adhésion à l'Union est avant tout une démarche volontaire.

MICRO TROTTOIR 6 :	
J'ai l'impression que les décisions sont prises à un trop haut niveau. *On est trop loin du peuple.*	DE
Bruxelles est à cent lieues de ce qui se passe dans la vie quotidienne.	EN
C'est un peu, peut-être, technocrate et un peu loin des gens.	FR
Et on a pas vraiment grand chose à dire, hein.	FR
Les droits de l'individu. Je pense aux droits et aux libertés de chacun. *Les Etats doivent en tenir compte.*	SV
Les politiques se méfient du citoyen. *C'est pour cela qu'ils préfèrent limiter les référendums.*	DE

L'Europe écoute-t-elle les préoccupations de ses citoyens ?
Comment ceux-ci peuvent-ils encore se faire entendre des Institutions dans une Europe à 25 ?
Le citoyen se pose parfois la question et l'Europe lui semble parfois bien lointaine.
La Constitution introduit des mécanismes qui rendent l'Union plus démocratique. Exemples :
Si un tiers des **Parlements nationaux** juge qu'une proposition de la Commission empiète sur les compétences nationales, la Commission sera invitée à revoir sa copie.

D'un autre côté, les séances du **Conseil des Ministres** seront publiques dès lors qu'on y débattra d'une proposition législative de la Commission.
Mais la mesure la plus spectaculaire est sans doute l'introduction de **l'initiative citoyenne**.
En réunissant 1 million de signatures dans un nombre significatif d'Etats membres, les citoyens pourront inviter la Commission à proposer une législation dans un domaine particulier.
L'Europe n'est donc plus seulement une démocratie représentative, elle devient une démocratie participative.

CONCLUSION :

La Constitution doit aujourd'hui être ratifiée par les 25 Etats membres.
Elle est l'occasion d'un pas supplémentaire dans la construction européenne. Aux citoyens ou à leurs représentants de décider s'ils le franchissent.

2. The English script

Why do we need a European Constitution?

Rome, 29th October, 2004.
The 25 member states sign the treaty establishing the European Constitution.
This document which modernises the Union and the way its institutions work is now the subject of intense debate.
Many Europeans are concerned about the real significance and consequences of what is destined to become their common constitution. But what do they think and what do they expect from the Constitution?

VOX POP 1:	
There's a lack of principles.	FR
Let's stop thinking that Europe is just one big market.	FR
It would be good if Europe could ensure peace and prosperity in all Member States.	FR
We need a European constitution to know what awaits us and what it's all about	SK
There should be a constitution otherwise the whole thing becomes a headless system.	DE
I don't think that it will damage the Slovakian constitution.	SK

It all started with coal and steel.
Then, step by step, the European Union was built.
Today, the "machine" is up and running.
But it is increasingly complex – a real maze.
It's difficult to understand all the ins and outs and Europe's «user manual» is becoming bigger and bigger.
Today, the Union is governed by 8 treaties.
They've piled up over time: from Rome to Nice, via Maastricht, the Single European act, Amsterdam …
The Constitution replaces all of them. Simpler and clearer, it has become **the single foundation** of the Union.
The text starts out by defining the Union's **objectives**:

live peacefully in freedom, security and justice; to create a stimulating economic environment; to promote research, sustainable development and to ensure a good standard of living for everyone.
The Constitution obliges the members of the Union to respect **values.** Equal opportunities, tolerance, solidarity, minority rights …
It goes on to define **European citizenship.**
It guarantees additional liberties, such as freedom to travel and to live anywhere in the European Union.
This is something from which people from every member state benefit, in addition to their own national citizenship.
The Constitution also includes the **Union's Charter of Fundamental Rights.**
Civil, social, political and religious – even administrative or bio-ethical – rights enjoyed by everyone living in a Union member state.

VOX POP 2:	
Decision procedures, these are one black box for the ordinary people like me.	EN
It's not clear who does what.	DE
Our representatives in Brussels should tell us what it entails.	SK
It's grown so much that the rules that are now there aren't suitable	EN
I wonder whether the balance of power is correct.	DE
I feel that some states have more power than others. *The constitution will regulate how we decide together.*	SK
If the 25 Member States have the same power, it won't work. *They should rise above themselves and think in terms of Europe*	FR

The European ensemble has grown considerably in size.
Achieving harmony from a group such as this requires clear, precise rules.
Without them, there's a risk of false notes and blockages in decision making.
So the Constitution clarifies the role and the composition of the European institutions and modifies some operating rules.
First, **The European Council.**
Through the Constitution, it becomes an Institution in its own right.
Every three months, heads of state and government meet to agree on the broad political orientation of the Union.
With this orientation in mind, it's then up to the **Commission** to take concrete initiatives. It is the Commission, completely independently, which drafts European legislation, looking after the Union's interests.
There is currently one Commissioner for each member state.
To guarantee that the Commission can operate efficiently, in 2014, the date by which new countries should have joined the Union, the Constitution will limit the number of Commissioners to two-thirds of the number of member states, selected in rotation.
The Council of Ministers is there to represent the interests of the member states and adopts, or rejects, the Commission proposals.
Each country assigns an appropriate minister, depending on the subject matter.
With 25 members, unanimous voting becomes difficult.

Except for certain sensitive areas – taxation, for example – the Constitution applies the principle of qualified majority.
For a vote to be carried, at least 55% of the votes of Council members are needed, representing 65% of the population. This formula gives a double legitimacy to Union decisions, based on majorities of states and people.
Democratically elected by the people every 5 years, **The European Parliament** represents citizens' interests.
The Constitution limits the number of MEPs to 750, with a minimum of 6 members for the smaller countries and a maximum of 96 for the larger ones.
But, above all, the Constitution reinforces the legislative role of the Parliament, alongside the Council of Ministers.
Practically every project proposed by the Commission now has to have approval from the Parliament before it can be adopted by the Council.
As for the **Court of Justice**, it's there to ensure that everyone complies with European legislation, and this includes the European Union Institutions.
European legislation includes several dozen different types of legislative act. To make things more straightforward, the Constitution reduces this number to 6: ranging from a European law to a simple recommendation.

VOX POP 3:	
I think more power should be given to the European Union.	IT
We need a Europe of the people, a social Europe.	FR
Unemployment... Justice... Fiscal pressure... *It should be harmonised in Europe.*	*SV*
Nobody can really do what they want.	FR
That's the paradox. Unity is important, but also individuality.	FR
Law after law, regulation after regulation	EN
Everything is controlled, they are control freaks	EN

So will the European Union simply be a servant of market forces…
A predator, hungry for power…
A threat to the sovereignty of its members…?
Three principles in the Constitution prohibit this.
Firstly: The Union can act only in the areas expressly assigned to it by member states.
Secondly: The Union can intervene only if its action provides added value to actions taken by member states individually.
Finally: The Union can not exceed that which is strictly necessary to achieve the stated objectives.
But what, precisely, are the Union's competencies?
ONE: « **exclusive competencies** ». Monetary policy concerning the euro and competition rules in the context of the internal market for example. These are areas where only the European Union may act.
TWO: « shared **competencies** » such as: agriculture, fisheries, energy, transport and environmental protection.
Those domains in which the Union AND the member states, can take complementary initiatives.

THREE: In areas such as health, education, culture and industrial policy, national policies are still preferred, but the Union can, nevertheless, give **support** or **coordinate** action.

VOX POP 4:	
We should have a European foreign policy.	FR
I think they should have a common defence policy	EN
Self-interest will have to give way. The world is one big village.	DE
In the case of Kosovo, Europe also takes action as one power. They can prove there whether it's worth it or not.	DE
I think that a Minister of Foreign Affairs at European level for the whole Union would be a good idea.	IT
One reference person should be appointed to represent Europe.	FR

Events have shown that the European Union needs to increase its influence and acquire more weight on the international stage.
Henry Kissinger once joked: « Europe? Sure, what phone number?»
The Constitution now provides him with a clear answer.
It creates the new post of **European Minister for Foreign Affairs**:
Appointed for a 5 year period by the European Council, he or she will be the voice and face of the Union to the outside world.
The new minister will carry out the Union's Foreign and Security policies.
The member states will retain their prerogative in these areas, but the minister will try to establish common positions.
Another innovation, which will strengthen Europe's role on the world stage: the Constitution endows the Union with a legal persona.
This is vital, most notably to be able to have a seat representing Europe as a whole on international bodies.
One day we could see Europe preside at the UN Security Council in the name of all its member states.

VOX POP 5:	
I think that my son will have more opportunities. In that sense I am definitely an optimist.	SK
They should start with a simpler system for students to study elsewhere.	SV
Certain values are lost, as is the personal character.	FR
The stupid laws, the way it's taking away our independence.	EN
I like the idea of the same money	EN
…the records on human rights	EN
Cheese producers can't sell their products. Factories close down because of EU requirements. It's scandalous.	FR
– If they want to leave, let them.	FR
It's a dictatorship after all, you know, I don't want to live in a dictatorship.	EN
Either you're in or you're out. You can't do both, can you.	EN

The Union is open to all European states which share its values and its objectives..
But it does not intend to be a ball and chain.
So, the Constitution provides for the possibility of withdrawing from the Union.
This clause is a reminder that joining the Union is voluntary.

VOX POP 6:	
I feel that decisions are taken at too high a level. It's too far away from the people.	DE
Brussels is so far away and it has no connections to people's everyday lives.	EN
It's a bit technocratic and too far removed from the people.	FR
We don't have much say.	FR
The right of the individual. I think of the rights and the liberties of each. The states need to take them into account.	SV
The political world doesn't trust citizens. *And that's why they don't want too many referendums.*	DE

Does Europe listen to its people?
How can individuals make themselves heard by the institutions in a Europe of 25?
For people who ask themselves this sort of question, Europe often seems a long way away.
The Constitution introduces mechanisms to make the Union more democratic.

Some examples:
If one third of **national parliaments** judge that a Commission proposal impinges upon areas of national competence, the Commission will be asked to review it.
Additionally, **Council of Ministers** meetings will be open to the public when a legislative proposal from the Commission is being debated.
But the most impressive measure must be the introduction of the **Citizen's Initiative.**
By collecting one million signatures in a certain number of member states, citizens can ask the Commission to propose legislation in particular areas.
Europe therefore is no longer just a representative democracy; it's a democracy the people can take part in too.

CONCLUSION:

The Constitution now has to be ratified by the 25 member states.
It's one more step in the construction of Europe. Now it is up to the people – or their representatives – to decide whether it's a step they want to take.

3. The Italian script

Perché una Costituzione europea?

Roma, 29 ottobre 2004.
I 25 firmano il trattato che a istituisce la Costituzione per l'Europa.
Questo testo, che modernizza l'Unione e il funzionamento delle sue istituzioni, è oggi al centro di un intenso dibattito.
Molti europei s'interrogano infatti sul significato reale e sulle conseguenze di quella che è destinata ad essere la loro Costituzione comune. Ma che cosa pensano dell'Unione europea e che cosa si aspettano dalla sua Costituzione?

VOX POP 1:	
Ci mancano i grandi principi.	FR
Smettere di pensare che l'Europa sia solo un grande mercato.	FR
Sarebbe una gran cosa se l'Europa portasse pace e crescita a tutti i suoi membri.	FR
Abbiamo bisogno di una costituzione europea per sapere ciò che ci attende e di cosa si tratta.	SK
Ci vuole una costituzione, altrimenti l'Europa sarebbe una casa senza tetto.	DE
Non credo che la costituzione europea sminuisca la costituzione slovacca.	SK

Tutto cominciò con il carbone e l'acciaio.
Poi, tappa dopo tappa, è stata costruita l'Unione europea.
Oggi la «macchina» funziona.
Ma è sempre più complessa: un vero dedalo!
È difficile comprenderne tutti i meandri. La "guida all'uso" dell'Europa è sempre più corposa.

Oggi l'Unione si fonda su 8 trattati.
Si sono accumulati nel corso del tempo: da Roma a Nizza, passando per Maastricht, l'Atto unico, Amsterdam...
La Costituzione li sostituisce tutti. Più semplice e più chiara, essa diventa la **base unica** dell'Unione.
Per cominciare, il testo definisce gli **obiettivi** dell'Unione:
vivere in pace in uno spazio di libertà, sicurezza e giustizia, creare un ambiente economico stimolante, promuovere la ricerca, lo sviluppo sostenibile e assicurare un buon tenore di vita per tutti.
La Costituzione vincola l'appartenenza all'Unione al rispetto di **valori.** Pari opportunità, tolleranza, solidarietà, rispetto delle minoranze…
Essa istituisce inoltre la **cittadinanza dell'Unione.**
Garantisce libertà supplementari, come quella di circolare e soggiornare liberamente in qualsiasi paese dell'Unione europea.
Tutti i cittadini degli Stati membri ne beneficiano, in aggiunta alla cittadinanza nazionale.
La Costituzione riprende infine la **Carta dei diritti fondamentali dell'Unione.**
Diritti civili, sociali, politici e religiosi, e persino diritti amministrativi o bioetici, di cui godono tutte le persone che vivono nel territorio dell'Unione.

VOX POP 2:	
Il processo decisionale è la bestia nera di noi persone comuni.	EN
Non si sa davvero chi fa che cosa.	DE
I nostri rappresentanti a Bruxelles: sta a loro dirci come votano e perché.	SK
L'Europa è talmente cresciuta che le regole non vi si adattano più.	EN
Mi chiedo se i rapporti di forza siano equilibrati.	DE
Mi pare che certi stati abbiano più peso di altri. *Grazie alla costituzione decideremo meglio tutti insieme.*	SK
Non si può dare a tutti e 25 lo stesso peso politico. Bisogna "pensare Europa" e andare oltre l'idea di nazione.	FR

La formazione europea si è notevolmente ingrandita.
Per accordare un tale ensemble sono necessarie regole chiare e precise, senza le quali si rischiano le stonature e il blocco di ogni decisione.
La Costituzione chiarisce quindi il ruolo e la composizione delle istituzioni europee e modifica alcune regole di funzionamento.
Innanzitutto, **il Consiglio europeo**.
Con la Costituzione, esso diventa un'istituzione a pieno titolo.
Ogni tre mesi, i capi di Stato e di governo si riuniscono per definire gli orientamenti politici generali dell'Unione.
In funzione di tali orientamenti, spetta quindi alla **Commissione** prendere iniziative concrete.
Essa elabora i progetti di leggi europee in piena indipendenza e nell'interesse dell'Unione.
Attualmente è composta da un Commissario per ogni Stato membro.
Per garantire il buon funzionamento della Commissione, dal 2014, data alla quale nuovi paesi dovrebbero aver aderito all'Unione, la Costituzione limiterà il numero di Commissari a 2/3 del numero di Stati membri, scelti sulla base di una rotazione.
Il Consiglio dei ministri rappresenta gli interessi degli Stati membri. Adotta o respinge i progetti proposti dalla Commissione.
In funzione della materia trattata, ogni paese delega il proprio ministro competente.
Con 25 paesi, il voto all'unanimità diventa difficile.
A parte alcuni settori delicati – per esempio, la fiscalità – la Costituzione generalizza la regola della maggioranza qualificata.
Per raggiungerla, sono necessari i voti di almeno il 55% dei membri del Consiglio, che rappresentano il 65% della popolazione. Questa formula conferisce all'Unione una doppia legittimità, fondata sulla maggioranza degli Stati e della popolazione.
Eletto a suffragio universale ogni 5 anni, **il Parlamento europeo** rappresenta gli interessi dei cittadini.
La Costituzione limita il numero di deputati a 750, con un minimo di 6 seggi per i paesi più piccoli e un massimo di 96 per quelli più grandi.
Tuttavia, la Costituzione rafforza soprattutto il ruolo di legislatore del Parlamento, al fianco del Consiglio dei ministri.
Quasi nessun progetto presentato dalla Commissione potrà essere adottato dal Consiglio senza l'approvazione del Parlamento.
La **Corte di giustizia** assicura il rispetto del diritto europeo da parte di tutti, comprese le istituzioni dell'Unione europea.
Il diritto europeo conta decine di diversi tipi di atti legislativi. A fini di chiarezza, la Costituzione ne riduce il numero a 6: dalle leggi europee alle semplici raccomandazioni.

VOX POP 3:	
Penso che sia importante dare ancora di più potere all'Unione europea.	IT
L'Europa dei popoli e l'Europa sociale sono ancora da costruire.	FR
L'occupazione. La giustizia. Il fisco. Vanno armonizzati in tutta Europa.	SV
Non si può più fare ciò che si vuole a casa propria.	FR
Il paradosso, è fare l'unità conservando i propri costumi.	FR
Legge su legge, regola su regola.	EN
Tutto controllato. Sono dei controllori nati.	EN

Quindi l'Europa sarebbe solo al servizio delle forze di mercato…
Un predatore, avido di potere…
Una minaccia per la sovranità dei suoi membri...?
Tre principi sanciti dalla Costituzione lo impediscono.
In primo luogo: l'Unione può agire solo nei settori che le sono espressamente affidati dagli Stati membri.
In secondo luogo: l'Unione può intervenire solo se la sua azione apporta un valore aggiunto rispetto a quelle intraprese individualmente dagli Stati membri.
Infine: l'Unione non può andare al di là di quanto strettamente necessario per il conseguimento degli obiettivi fissati.
Ma quali sono esattamente le competenze dell'Unione?
UNO: «**competenze esclusive**». La politica monetaria relativa all'euro e le norme che disciplinano il funzionamento del mercato interno, per esempio. Si tratta di settori nei quali può agire solo l'Unione europea.
DUE: «**competenze condivise**», per esempio: l'agricoltura, la pesca, l'energia, i trasporti, la tutela dell'ambiente.
Si tratta dei settori nei quali l'Unione e gli Stati membri possono prendere iniziative complementari.
TRE: In settori quali la sanità, l'istruzione, la cultura e la politica industriale si continuano a preferire le politiche nazionali; l'Unione può tuttavia condurre azioni **di sostegno** o di **coordinamento**.

VOX POP 4:	
Ci vuole una vera politica estera europea.	FR
Ci vuole una politica di difesa comune.	EN
L'interesse nazionale deve sparire, il mondo è ormai un solo villaggio.	DE
Riguardo al Kosovo, l'Europa agisce a una sola voce. *E qui che possono dimostrare che serve a qualcosa.*	DE
Penso che avere un Ministro degli Affari Esteri al livello Europeo che possa decidere per tutta quanta la Comunità sia un'ottima idea.	IT
Ci vorrebbe una persona di prestigio per rappresentare l'Europa.	FR

I fatti dimostrano che l'Unione europea deve accrescere la sua influenza e acquisire maggior peso sulla scena internazionale.
«Europe? Sure, what phone number?» ironizzava un giorno Henry Kissinger.
La Costituzione fornisce ora una risposta chiara.
Essa istituisce infatti la figura del **ministro europeo degli Affari esteri.**
Nominato per un periodo di 5 anni dal Consiglio europeo, sarà la voce e il volto dell'Unione nei confronti del mondo esterno.
Condurrà la politica estera e di sicurezza comune dell'Unione.
Gli Stati membri conservano le loro prerogative in questi settori, ma il ministro cercherà di definire posizioni comuni.
Un'altra innovazione rafforzerà il ruolo dell'Europa nel mondo: la Costituzione conferisce all'Unione personalità giuridica.
Ciò è indispensabile, in particolare per poter sedere negli organismi internazionali e rappresentare l'Europa nel suo insieme.

Un giorno potremo quindi vedere l'Unione sedere in seno al Consiglio di sicurezza delle Nazioni Unite a nome di tutti i suoi Stati membri.

VOX POP 5:	
Penso che mio figlio avrà più possibilità di me. Sono ottimista a questo riguardo.	SK
Si dovrebbe incominciare semplificando il sistema di scambio degli studenti.	SV
Non si trovano più i propri valori. Si perde un po' della propria personalità.	FR
Fanno delle leggi stupide. Ci tolgono l'indipendenza.	EN
Amo l'idea della moneta unica.	EN
E il progresso dei diritti dell'uomo.	EN
Le fabbriche non a norma europea devono chiudere. È uno scandalo.	FR
Se vogliono andarsene, che vadano.	FR
È una dittatura. Non intendo vivere in una dittatura.	EN
Non si può stare con un piede dentro e uno fuori.	EN

L'Unione è aperta a tutti gli Stati europei che condividono i suoi valori e i suoi obiettivi, ma non intende diventare una palla al piede.
La Costituzione prevede quindi la possibilità di ritirarsi dall'Unione.
Questa clausola rammenta che l'adesione all'Unione è una scelta volontaria.
Per contro, la Costituzione prevede le «cooperazioni rafforzate» per i paesi che intendono procedere più rapidamente in alcuni settori. Ciò avviene già per quanto riguarda la soppressione dei controlli alle frontiere interne o nell'ambito della cooperazione di polizia e giudiziaria.
L'idea è di creare un'avanguardia cui si possano unire gli altri paesi in un secondo tempo.

VOX POP 6:	
Ho l'impressione che le decisioni siano prese lassù molto lontano dalla gente.	DE
Bruxelles è a mille miglia da quella che è la vita quotidiana.	EN
E un mix di tecnocrazia e di distacco dalla gente.	FR
Non ci resta più molto da dire.	FR
I diritti dell'individuo. Penso ai diritti e alle libertà di ciascuno. Gli Stati devono ternerne conto.	SV
I politici non si fidano dei cittadini. Per questo preferiscono limitare i referendum.	DE

L'Europa ascolta i suoi cittadini?
Come si può far sentire la propria voce alle istituzioni di un'Europa a 25?
Al cittadino che si pone questi interrogativi, l'Europa sembra spesso molto lontana.
La Costituzione introduce meccanismi che rendono l'Unione più democratica.

Esempi:
Se un terzo dei **parlamenti nazionali** ritiene che una proposta della Commissione interferisca con le competenze nazionali, la Commissione sarà invitata a riesaminarla.
Inoltre, le sedute del **Consiglio dei ministri** in cui si discutono le proposte legislative della Commissione saranno pubbliche.

Tuttavia, la misura più spettacolare è senza dubbio l'introduzione dell'**iniziativa popolare**.
Raccogliendo un milione di firme in un determinato numero di Stati membri, i cittadini possono indurre la Commissione a presentare proposte legislative in determinati settori.
Pertanto, l'Europa non è più solo una democrazia rappresentativa, ma diventa una democrazia partecipativa.

CONCLUSIONE:

La Costituzione deve ora essere ratificata dai 25 Stati membri.
È un passo supplementare nella costruzione europea. Spetta ora ai cittadini o ai loro rappresentanti decidere se compiere tale passo.

Name index

Subject index

In the series *IMPACT: Studies in language and society* the following titles have been published thus far or are scheduled for publication:

28 **NORDE, Muriel, Bob de JONGE and Cornelius HASSELBLATT (eds.):** Language Contact. New perspectives. 2010. vii, 225 pp.

27 **EDWARDS, John:** Minority Languages and Group Identity. Cases and Categories. 2010. ix, 231 pp.

26 **BEECHING, Kate, Nigel ARMSTRONG and Françoise GADET (eds.):** Sociolinguistic Variation in Contemporary French. 2009. xi, 257 pp.

25 **STANFORD, James N. and Dennis R. PRESTON (eds.):** Variation in Indigenous Minority Languages. 2009. vii, 519 pp.

24 **MEYERHOFF, Miriam and Naomi NAGY (eds.):** Social Lives in Language – Sociolinguistics and multilingual speech communities. Celebrating the work of Gillian Sankoff. 2008. ix, 365 pp.

23 **LEVEY, David:** Language Change and Variation in Gibraltar. 2008. xxii, 192 pp.

22 **POTOWSKI, Kim and Richard CAMERON (eds.):** Spanish in Contact. Policy, Social and Linguistic Inquiries. 2007. xx, 398 pp.

21 **HUTCHBY, Ian:** The Discourse of Child Counselling. 2007. xii, 145 pp.

20 **FENYVESI, Anna (ed.):** Hungarian Language Contact Outside Hungary. Studies on Hungarian as a minority language. 2005. xxii, 425 pp.

19 **DEUMERT, Ana:** Language Standardization and Language Change. The dynamics of Cape Dutch. 2004. xx, 362 pp.

18 **DEUMERT, Ana and Wim VANDENBUSSCHE (eds.):** Germanic Standardizations. Past to Present. 2003. vi, 480 pp.

17 **TRINCH, Shonna L.:** Latinas' Narratives of Domestic Abuse. Discrepant versions of violence. 2003. x, 315 pp.

16 **BRITAIN, David and Jenny CHESHIRE (eds.):** Social Dialectology. In honour of Peter Trudgill. 2003. x, 344 pp.

15 **BOXER, Diana:** Applying Sociolinguistics. Domains and face-to-face interaction. 2002. xii, 245 pp.

14 **WEBB, Victor:** Language in South Africa. The role of language in national transformation, reconstruction and development. 2002. xxviii, 357 pp.

13 **OAKES, Leigh:** Language and National Identity. Comparing France and Sweden. 2001. x, 305 pp.

12 **OKITA, Toshie:** Invisible Work. Bilingualism, language choice and childrearing in intermarried families. 2002. x, 275 pp.

11 **HELLINGER, Marlis and Hadumod BUSSMANN (eds.):** Gender Across Languages. The linguistic representation of women and men. Volume 3. 2003. xiv, 391 pp.

10 **HELLINGER, Marlis and Hadumod BUSSMANN (eds.):** Gender Across Languages. The linguistic representation of women and men. Volume 2. 2002. xiv, 349 pp.

9 **HELLINGER, Marlis and Hadumod BUSSMANN (eds.):** Gender Across Languages. The linguistic representation of women and men. Volume 1. 2001. xiv, 329 pp.

8 **ARMSTRONG, Nigel:** Social and Stylistic Variation in Spoken French. A comparative approach. 2001. x, 278 pp.

7 **McCAFFERTY, Kevin:** Ethnicity and Language Change. English in (London)Derry, Northern Ireland. 2001. xx, 244 pp.

6 **RICENTO, Thomas (ed.):** Ideology, Politics and Language Policies. Focus on English. 2000. x, 197 pp.

5 **ANDREWS, David R.:** Sociocultural Perspectives on Language Change in Diaspora. Soviet immigrants in the United States. 1999. xviii, 182 pp.

4 **OWENS, Jonathan:** Neighborhood and Ancestry. Variation in the spoken Arabic of Maiduguri, Nigeria. 1998. xiv, 390 pp.

3 **LINELL, Per:** Approaching Dialogue. Talk, interaction and contexts in dialogical perspectives. 1998. xvii, 330 pp.

2 **KIBBEE, Douglas A. (ed.):** Language Legislation and Linguistic Rights. Selected Proceedings of the Language Legislation and Linguistic Rights Conference, the University of Illinois at Urbana-Champaign, March, 1996. 1998. xvi, 415 pp.

1 **PÜTZ, Martin (ed.):** Language Choices. Conditions, constraints, and consequences. 1997. xxi, 430 pp.