Dirk Geeraerts, Gitte Kristiansen,
Yves Peirsman (Eds.)

ADVANCES IN COGNITIVE SOCIOLINGUISTICS

Advances in Cognitive Sociolinguistics

Cognitive Linguistics Research 45

Editors
Dirk Geeraerts
John R. Taylor

Honorary editors
René Dirven
Ronald W. Langacker

De Gruyter Mouton

Advances in Cognitive Sociolinguistics

Edited by
Dirk Geeraerts
Gitte Kristiansen
Yves Peirsman

De Gruyter Mouton

Table of contents

Part three: Variation of lectal awareness and attitudes

List of contributors

Raphael Berthele
University of Freiburg
Institut für Mehrsprachigkeit
Murtengasse 24
1700 Freiburg
Switzerland
raphael.berthele@unifr.ch

Lynn Clark
Lancaster University
Department of Linguistics and
English Language
Lancaster LA1 4YT
United Kingdom
l.clark6@lancaster.ac.uk

Timothy Colleman
Ghent University
Department of Dutch Linguistics
Blandijnberg 2
9000 Gent
Belgium
timothy.colleman@ugent.be

Gunther De Vogelaer
Research Foundation Flanders
(FWO) / Ghent University
Department of Dutch linguistics
Blandijnberg 2
9000 Gent
Belgium
gunther.devogelaer@ugent.be

Dirk Geeraerts
University of Leuven
Department of Linguistics
Blijde-Inkomststraat 21
P.O.Box 3308
3000 Leuven
Belgium
dirk.geeraerts@arts.kuleuven.be

Kris Heylen
University of Leuven
Department of Linguistics
Blijde-Inkomststraat 21
P.O.Box 3308
3000 Leuven
Belgium
kris.heylen@arts.kuleuven.be

Gitte Kristiansen
English Language and Linguistics
Facultad de Filologia
Universidad Complutense de
Madrid
Ciudad Universitaria
28040 Madrid
Spain
gkristia@filol.ucm.es

Yves Peirsman
University of Leuven
Department of Linguistics
Blijde-Inkomststraat 21
P.O.Box 3308
3000 Leuven
Belgium
yves.peirsman@arts.kuleuven.be

Justyna Robinson
University of Leeds
School of English
Woodhouse Lane
Leeds LS2 9JT
United Kingdom
j.a.robinson@leeds.ac.uk

Augusto Soares da Silva
Portuguese Catholic University
Faculdade de Filosofia
P-4710-297 Braga
Portugal
assilva@braga.ucp.pt

Dirk Speelman
University of Leuven
Department of Linguistics
Blijde-Inkomststraat 21
P.O.Box 3308
3000 Leuven
Belgium
dirk.speelman@arts.kuleuven.be

Benedikt Szmrecsanyi
Freiburg Institute for Advanced
Studies
Albertstr. 19
79104 Freiburg
Germany
bszm@frias.uni-freiburg.de

Graeme Trousdale
University of Edinburgh
Linguistics and English Lan-
guage
Dugald Stewart Building
3 Charles Street
Edinburgh EH8 9AD
United Kingdom
graeme.trousdale@ed.ac.uk

Introduction.
Advances in Cognitive Sociolinguistics

Dirk Geeraerts, Gitte Kristiansen and Yves Peirsman

Cognitive Sociolinguistics is a novel and burgeoning field of research which seeks to foster investigation into the socio-cognitive dimensions of language at a usage-based level. The ten chapters compiled in the present volume grew out of presentations at the Theme Session "Cognitive Sociolinguistics", celebrated at the 10th International Cognitive Linguistics Conference in Kraków, Poland. While the first volume to bear the title *Cognitive Sociolinguistics* (Kristiansen and Dirven 2008) took a fairly broad view on the field by exemplifying a variety of areas where sociolinguistics and Cognitive Linguistics meet in natural manners, this volume focuses specifically on language-internal variation and methodological advances in the field. Emphasis is on empirical usage-based variation research and as a natural corollary also on the methods of cognitive sociolinguistic inquiry.

The ten studies brought together in *Advances in Cognitive Sociolinguistics*, then, are tied together by theoretical, methodological, and descriptive characteristics. Theoretically speaking, *Advances in Cognitive Sociolinguistics* brings together ten studies into the social and conceptual aspects of language-internal variation. Methodologically speaking, all ten contributions rely on a firm empirical basis in the form of advanced corpus-based techniques, experimental methods and survey-based research, or a combination of these. The search for methods that may adequately unravel the complex and multivariate dimensions intervening in the interplay between conceptual meaning and variationist factors is thus another characteristic of the volume. Finally, in terms of its descriptive scope, the volume covers three main areas: lexical and lexical-semantic variation, constructional variation, and research on lectal attitudes and acquisition. It thus illustrates how Cognitive Sociolinguistics studies both the variation of meaning, and the meaning of variation.

In this Introduction, we will first introduce Cognitive Sociolinguistics as a convergence of the concerns of sociolinguistics and Cognitive Linguistics. In particular, we will try to indicate how both traditions may profit from a confrontation. The second part of the Introduction summarizes the

various contributions to the volume and positions them against the background of the scope of Cognitive Sociolinguistics as it emerges from the first section. Given the series in which this book appears, the perspective that we will take in this Introduction is specifically targeting scholars with a background in Cognitive Linguistics. We envisage a mutually beneficial interaction of both approaches, but in describing that bidirectionality, we will start from the Cognitive Linguistics end. (Note also that we capitalize 'Cognitive Linguistics' because we think of it as a specific theoretical framework – even if it is a multidimensional one – whereas uncapitalized 'sociolinguistics' refers primarily to a domain of research rather than a specific theoretical outlook.)

1.   The nature and scope of Cognitive Sociolinguistics

Cognitive Sociolinguistics may be broadly defined as the attempt to achieve a convergence of Cognitive Linguistics and the tradition of sociolinguistics. Two questions then arise: why would Cognitive Linguistics turn to variationist research, and why would sociolinguistics bother about Cognitive Linguistics? Let us try to answer both questions in brief.

1.1. The social perspective in Cognitive Linguistics

Reasoning from the perspective of Cognitive Linguistics, there are two defining aspects of the approach that lead towards the incorporation of social variation: the predominantly *semantic* perspective of Cognitive Linguistics, and the *usage-based* nature of Cognitive Linguistics. Both aspects (which are themselves interrelated in various ways) are defining features to the extent that they lie at the heart of the cognitive linguistic enterprise, and to a large extent determine the internal development of the approach. Presenting each of the two features in detail is beyond the scope of the present introduction, but a few references may help to bring the points to mind. For each of the features, we will specify how they are inevitably linked up with a social perspective.

1. It hardly needs to be spelled out that the study of linguistic meaning constitutes the foundational characteristic par excellence of Cognitive Linguistics. In Geeraerts (2006), for instance, written as an introduction to a collec-

tion of basic readings in Cognitive Linguistics, it is shown how a focus on meaning constitutes a basic principle of cohesion for Cognitive Linguistics. The central concepts of Cognitive Linguistics are mutually related because they derive from a common set of underlying fundamentals: the principle that language is all about meaning, in the broadest possible sense, in combination with four specific assumptions about the nature of linguistic meaning – that meaning is flexible and dynamic, that it is encyclopedic and non-autonomous, that it is based on usage and experience, and that it is perspectival in nature.

But meaning does not exist in isolation: it is created in and transmitted through the interaction of people, and that is why the definition and the basic architecture of language are recognized by Cognitive Linguistics as involving not just cognition, but socially and culturally situated cognition. A specific line of studies produced in the context of Cognitive Linguistics analyzes the way in which the emergence of language as such and the presence of specific features in a language can only be adequately conceived of if one takes into account the socially interactive nature of linguistic communication. Examples of this strand of research include Sinha (2007) on language as an epigenetic system, Zlatev (2005) on situated embodiment, Itkonen (2003) on the social nature of the linguistic system, Verhagen (2005) on the central role of intersubjectivity in language, and Harder (2003) on the socio-functional background of language.

The references mentioned here mostly take a foundational rather than a descriptive point of view: establishing the social nature of meaning as such predominates over the empirical study of variation as illustrated in the present volume. It needs to be pointed out, though, that such foundational studies are important, because the social nature of meaning has been less spontaneously obvious to Cognitive Linguistics than we suggested above. In fact, although the notion of *cultural model* played a significant role in the emergence of Cognitive Linguistics as a linguistic new framework (Holland and Quinn 1987), a certain tension exists within Cognitive Linguistics between a more universalist approach and a more culturally oriented approach. A typical case in point is the discussion between Geeraerts and Grondelaers (1995) on the one hand and Kövecses (1995) on the other regarding the nature of ANGER IS HEAT metaphors: while the former emphasized the culturally specific and historically contingent nature of such metaphorical patterns, the latter defended a universalist, physiologically grounded position . In recent years, however, the socio-cultural pers-

pective has been gaining ground, even among those who previously opposed it: see Kövecses (2005).

2. There is a growing consensus within Cognitive Linguistics to conceive of itself as a usage-based approach to language. According to a number of programmatic accounts of usage-based linguistics (Langacker 1999; Kemmer and Barlow 2000; Bybee and Hopper 2001; Tomasello 2003; Verhagen and Van de Weijer 2003), the essential idea of a usage-based linguistics is the dialectic nature of the relation between language use and the language system. The grammar does not only constitute a knowledge repository to be employed in language use, but it is also itself the product of language use. The former perspective considers usage events as specific, actual instantiations of the language system. According to this view, one can gain insight into the language system by analyzing the usage events that instantiate it. This is a strong motivation for empirical research: the usage data constitute the empirical foundation from which general patterns can be abstracted. The latter perspective considers usage events as the empirical source of the system. From this point of view, usage events define and continuously redefine the language system in a dynamic way. As a result, every usage event may slightly redefine a person's internal language system.

The consequences of such a position are both thematic and methodological. Methodologically speaking, you cannot have a usage-based linguistics unless you study actual usage, as it appears in an online and elicited form in experimental settings or as it appears in its most natural form in corpora in the shape of spontaneous, non-elicited language data. While it would be an exaggeration to say that the level of empirical grounding illustrated in the present volume is the norm in Cognitive Linguistics, we can definitely see that the interest in corpus-based and experimental studies is growing (cp. Tummers, Heylen, and Geeraerts 2005).

Thematically speaking, a usage-based approach fosters interest in specific topics and fields of investigation. For instance, it follows from the dialectic relationship between structure and use that the analysis of linguistic change (Bybee 2007) is a natural domain of application for any usage-based approach. Similarly, interesting perspectives for cognitive stylistics and poetics, and for language acquisition research open up. The usage-based approach holds the promise of answering the acquisition problem that looms large in the Chomskyan delimitation of linguistics. In the work done by Tomasello and his group (2003), an alternative is presented for the Chomskyan genetic argument. These researchers develop a model of lan-

guage acquisition in which each successive stage is (co)determined by the actual knowledge and use of the child at a given stage, i.e. language acquisition is described as a series of step by step usage-based extensions of the child's grammar. The grammar, so to speak, emerges from the child's interactive performance.

Now, one major consequence of a usage-base conception of language is that it needs to incorporate socio-variationist studies. To see why this is the case, we need to have a closer look at the dialectic relationship between system and use that is the hallmark of the usage-based view. How, in such a dialectic view of the relationship between structure and use, does the system exist – if at all? The 'use' side of the dialectic relationship is readily identifiable: it exists in the form of actual instances of language use, whether active or passive. But where do we find 'structure'? Analytically, we may argue as follows. First, language as structure is a social fact, as an observable regularity in the language use realized by a specific community. Second, it is at the same time a cognitive fact, because the members of the community have an internal representation of the existing regularities (the system) that allows them to realize the same system in their own use of the language. Third, the same mechanism that allows the existing collective regularities to enter the individual minds is also the one that allows regularities to emerge to begin with, viz. mutual influence in social interaction. People influence each other's behavior, basically by co-operative imitation and adaptation, and in some cases by opposition and a desire for distinctiveness. Paying attention to what others do, however subconsciously, thus creates a mental representation of the collective tendencies in the behavior of the community; adapting one's own behavior to those tendencies, reaffirms and recreates the tendencies. And fourth, in the same way that the existing regularities emerged from actual interaction, changes may emerge; as such, a degree of variation is an inevitable aspect of any synchronic state of the language.

Crucially, however, these interactions that reproduce, perpetuate, and change the linguistic system do not occur between all the members of a linguistic community at the same time: individual language users do not interact with all the other members of a linguistic community, but only with a subset. In that respect, the communicative interactions are not only social events, they also reflect social structure – a structure formed by the social groups and networks that a communicating individual belongs to and that shape his or her communicative interactions. We cannot, therefore, simply assume that 'the linguistic system' is uniform. Rather, the very concept of a

usage-based theory of language leads to the recognition of language-internal variety, of the kind that has been studied by sociolinguistics since many decades.

In actual fact, although there is a growing interest within Cognitive Linguistics for language-internal variation (Kristiansen 2003; Geeraerts 2005; Kristiansen and Dirven 2008; Croft 2009), it remains an understudied area. Too often linguistic analyses or cross-linguistic comparisons are carried out at the level of 'a language' as such, disregarding rich and complex patterns of intralingual variation. Such a level of granularity ultimately amounts to that of a homogeneous and thus idealized speech community. Cognitive Linguistics, to the extent that it takes the claim that it is a usage-based approach to language and cognition seriously, cannot afford to work with language situated taxonomically at an almost Chomskyan level of abstraction.

1.2. The cognitive perspective in sociolinguistics

Given that there appear to be compelling reasons for Cognitive Linguistics to take language-internal variation seriously and to adopt a socio-variationist perspective, could an analogous argumentation be formulated from the point of view of sociolinguistics? If we start from the rich tradition of social studies in linguistics, what would be the most obvious contribution of Cognitive Linguistics to sociolinguistics? Precisely because socio-variationist studies within Cognitive Linguistics are as yet merely an emerging field, we are entering largely programmatic domain with this question. Still, the analysis that we presented in the previous section indicates in what direction the answer may be sought: if the study of meaning is the core business of Cognitive Linguistics, then that is exactly the domain where we need to look for innovation and inspiration flowing from Cognitive Linguistics to sociolinguistics. Elaborating on a phrase that we used in the opening passage of this Introduction, there are two aspects to that approach: on the one hand, we may study variation of meaning, and on the other, the meaning of variation. Let us try to specify both aspects.

1. A most natural question to ask for Cognitive Sociolinguistics is: how does language-internal variation affect the occurrence of linguistic phenomena that have the specific attention of Cognitive Linguistics, i.e. meaning. But as *variation of meaning* is a largely understudied field in sociolinguis-

tics, that alone might suffice to accept Cognitive Sociolinguistics as a welcome addition to sociolinguistics. The importance of meaning for sociolinguistics is even more fundamental, however, because questions of meaning implicitly lie at the heart of the sociolinguistic enterprise as a whole. This can be easily understood if we take into account that the standard methodology of socio-variationist research involves the concept of a 'sociolinguistic variable'. Put simply, a sociolinguistic variable in the sense of contemporary sociolinguistics is a set of alternative ways of expressing the same linguistic function or realizing the same linguistic element, where each of the alternatives has social significance: 'Social and stylistic variation presuppose the option of saying 'the same thing' in several different ways: that is, the variants are identical in reference or truth value, but opposed in their social and/or stylistic significance' (Labov 1972: 271). As such, a sociolinguistic variable is a linguistic element that is sensitive to a number of extralinguistic independent variables like social class, age, sex, geographical group location, ethnic group, or contextual style and register.

Given, then, that the very notion of a sociolinguistic variable refers to meaning, it is important for sociolinguistic studies to devote specific attention to semantics – all the more so since specific difficulties arise when dealing with meaning: the interaction of meaning with other sources of variation, and the problem of semantic equivalence. We may illustrate both problems by having a closer look at lexical meaning, but the problems that we mention for lexical meaning clearly also apply to grammatical and other types of meaning.

First, observe that lexical choices in discourse may be determined by different factors: next to (obviously) the topic of the text, there is variation of a sociolinguistic or stylistic nature: at least for a number of concepts, a speaker of British English, for instance, will make different lexical choices than a speaker of American English. Choices of this kind often involve differences between language varieties. Lexical choices of this type are not choices for specific concepts (like the topic-related choices would be), but they are choices for one word rather than the other expressing the same concept: we can recognize American English when we come across the word *subway* in contrast with British English *underground*, but the type of public transport referred to is the same. We may call this type of variation (i.e. the *subway/underground* type) *formal onomasiological variation (FOV)*, in contrast with *conceptual onomasiological variation (COV)*, which involves thematic choices, like talking about public transport rather than beer, biology, or Bach. For the third type of variation, we may use the

term *speaker and situation related variation (SSV)*. It covers all relevant features of the speech situation: not just the lectal variation that comes with more or less permanent speaker characteristics (like being American or British), but also the more transient, interactional characteristics of the communicative context, like whether the speech event is a dialogue or a monologue.

The interaction between these various dimensions has not yet been systematically investigated. Such interactions may surely be expected, and they are likely to work in different directions. The *underground/subway* case is an example of an SSV-FOV interaction, as would be choosing an informal term rather than a more formal one according to the formality of the speech situation. But at the same time, the choice for an informal expression might correlate with thematic factors: there might well be more dirty words for dirty topics than colloquial words for scientific topics. Whereas such a case would constitute a COV-FOV interaction, COV-SSV interactions may occur just as well. Conceptual choices, in fact, are not just determined by the topic of a text: for a number of specific concepts, they rather derive from the situational, interactional characteristics of the communicative context. Second person pronouns constitute an obvious example: they are likely to occur typically in dialogues rather than monologues. Similarly, persuasive texts contain different modal verbs than informative texts.

Given such interactions, the basic research question for sociolexicology (and more generally, for any semantically enriched type of variationist research) can be defined as follows: *what is the overall structure of lexical variation in terms of the relationship between FOV, COV, and SSV?* The specific situation of sociolexicological research in the context of sociolinguistics follows in a straightforward way from this question: while the paragon cases of sociolinguistic research involve formal variables and a binary relationship between formal variation and lectal context, sociolexicological research has to come to terms with a ternary relationship between form, meaning, and context.

The second problem we have to deal with when we introduce meaning into sociovariationist research is how to establish equivalence of meaning. Again, we may use the lexicon as an example. Treating onomasiological variation as a sociolinguistic variable means coming to terms with the meaning of words: the selection of a word is also the selection of a conceptual category, so if we are interested in the contextual choice between synonyms as an expression of sociolinguistic factors, we first need to control

for meaning – and that, needless to say, is not an obvious matter. The problem is methodological in a high-level sense: in what ways can we push semantic description beyond intuitive interpretation? But the problem is also methodological in a very practical sense: in what ways can we achieve a method of semantic interpretation – and more specifically, synonym identification – that works efficiently enough to allow for an easy demarcation of a large set of concepts? Getting a good grip on the interrelations between FOV, COV and SSV cannot be achieved unless we can study a sufficiently high number of concepts, but that ideally requires a method of semantic analysis that is as fast as it is trustworthy.

Within the field of sociolinguistics, the methodological problem of semantic equivalence was recognized early on by Beatriz Lavandera. She argued that 'it is inadequate at the current state of sociolinguistic research to extend to other levels of analysis of variation the notion of sociolinguistic variable originally developed on the basis of phonological data. The quantitative studies of variation which deal with morphological, syntactic, and lexical alternation suffer from the lack of an articulated theory of meanings' (Lavandera 1978: 171). In the mainstream development of sociolinguistics, however, the question of semantic equivalence, as a methodological prerequisite for the sociovariationist study of lexis and grammar, was not systematically pursued. What we may suggest as a preferred area of investigation for Cognitive Sociolinguistics, then, is to have a renewed look at Lavandera's question: for one thing, the question is still relevant within sociolinguistics, and for another, if there is one type of theoretical linguistics that has sufficient affinity with semantics to tackle the question, it must surely be Cognitive Linguistics – which does not equal saying that it has already solved the question.

2. As natural as it is for Cognitive Linguistics to study the variation of meaning, is it just as natural to study the *meaning of variation*, i.e. the way in which language users make sense of linguistic variation, the way in which linguistic variation is meaningful to them. In a usage-based conception of language, we assume that language users have a cognitive representation of the communicative interactions in which they participate: thát – rather than some genetic endowment – is their 'knowledge of the language'. But as their interactive horizon includes linguistic variation, they also have a representation of that diversity. They categorize social reality as reflected in language use and differences of language use, and such a categorization process is typically one of the phenomena that Cognitive Lin-

guistics is interested in: 'meaning as categorization' is a kind of catch-phrase in Cognitive Linguistics. The questions that arise here are of the following kind.

· How do language users perceive lectal differences, and how do they evaluate them attitudinally?
· What models do they use to categorize linguistic diversity?
· How does linguistic stereotyping work: how do language users categorize other groups of speakers?
· What is the role of subjective and objective linguistic distances: is there a correlation between objective linguistic distances, perceived distances, and language attitudes?
· Are there any cultural models of language diversity: what models of lectal variation, standardization, and language change do people work with?
· To what extent do attitudinal and perceptual factors have an influence on language change?
· How do language users acquire lectal competence, how is it stored mentally, and how does it work in language production?

From the point of view of the sociolinguistic tradition, this is the point where Cognitive Linguistics meets with perceptual dialectology, and to some extent with psycholinguistics. What Cognitive Linguistics can bring to this domain of investigation, are the various models of categorization (like prototypicality and cultural models) that it has developed in dealing with linguistic categories at large.

2.   Overview of the sections and contributions

Let us now have a closer look at the contents of the present volume, and see how it fits into the domain of Cognitive Sociolinguistics as defined in the previous pages. The volume is thematically structured in three sections. Part one comprises research on lexical and lexical-semantic dimensions of language-internal variation. Part two includes studies with an emphasis on grammatical and constructional aspects of lectal variation. The chapters in part three investigate attitudinal and acquisitional dimensions of varieties as such and of lectal-internal variables.

On a second level there is at the same time a methodological organization, which largely matches the divide between part one and part two on the one hand, and part three on the other. Most of the contributions in the first two sections implement advanced corpus-based techniques in order to deal with the multiple dimensions of language variation and to disentangle conceptual and social variation. In turn, the chapters in part three examine lectal varieties by means of experimental designs, surveys and questionnaires, showing how methods traditionally implemented in sociolinguistics and psycholinguistics can likewise throw light on the interaction between linguistic structure, cultural variation and conceptual content. Exceptions include the chapter in part two by De Vogelaer (who implements a questionnaire-based procedure) and the paper by Clark and Trousdale in part three (who combine a data-gathering process in the form of a variationist sociolinguistic survey with a multivariate statistical analysis of the data) – but these exceptions merely show that, regardless of the topic at hand, Cognitive Sociolinguistics enlists the aid of all empirical methods available to variationist researchers.

2.1. Part one. Lexical and lexical-semantic variation

This section comprises four chapters that in various manners examine lexical or lexical-semantic dimensions at the level of lectal variation. The first two contributions examine variation in lectal structure in relation to conceptual content and the last two papers apply novel methodological techniques to the study of semantic variation.

In "Heterodox concept features and onomasiological heterogeneity in dialects", **Dirk Geeraerts** and **Dirk Speelman** ask the question whether in dialectal variation not only geographical and social-stratificational factors contribute to lexical variation (as assumed in most dialectological and sociolinguistic research), but also the features of the concept itself: will non-traditional semasiological features such as conceptual salience and conceptual vagueness significantly influence the occurrence of onomasiological heterogeneity? The authors present the design and results of a study designed to test this hypothesis. Geeraerts and Speelman carried out a statistical analysis of a large-scale database with dialectological data for the Limburgish dialects of Dutch, operationalized the notions of vagueness, salience and negative effect, and performed a multiple linear regression analysis on the data. It is concluded that non-orthodox concept features

such as salience and vagueness have a marked effect on lexical hetero-
geneity and that an examination of dialectological lexical heterogeneity
purely from the point of view of geographic differentiation is too restric-
tive.

In the second contribution in this section, entitled "Measuring and pa-
rameterizing lexical convergence and divergence between European and
Brazilian Portuguese" **Augusto Soares da Silva** reports on a corpus-based
study into the lexical relationship between European Portuguese and Bra-
zilian Portuguese. In order to assess the degree of lexical divergence and
convergence between the two varieties in the course of the last 60 years and
throw light on the influence of internal linguistic parameters, Soares im-
plements quantitative methods designed to measure the *onomasiological
profile* (set of alternative synonymous terms used to designate a concept
together with their frequencies) and *uniformity* (similarity between the pro-
files of different varieties). The analysis was carried out for several thou-
sand observations of the use of alternative terms designating 43 nominal
concepts, gathered from large corpora: 21 sets of synonymous terms from
the lexical field of football and 22 profiles of clothing items. The author
concludes that both varieties diverge from each other in the vocabulary of
clothing, that the Brazilian variety has changed more than the European
variety and that the actual distance between the standard and the substan-
dard strata is higher in Brazilian Portuguese than in European Portuguese.

In "*Awesome* insights into semantic variation", **Justyna A. Robinson**
observes that in sociolinguistics few methods have been designed to deal
with social variation of meaning and that in cognitive semantics claims are
often made at an abstract level. Addressing the question whether variation-
ist sociolinguistics can provide insights into the conceptual structure of
polysemy and whether cognitive semantics can be of use for sociolinguis-
tics, Robinson combines cognitive and sociolinguistic analytical methods in
order to examine the flexibility of a polysemous category within the same
speech community. In her study, semasiological variants of the adjective
awesome were elicited in 72 one-to-one interviews and subjected to a cog-
nitive semantics analysis. As a second step, the various senses of *awesome*
were related to the sociolinguistic variables of age, gender, education, oc-
cupation and place of residence and subjected to a logistic regression analy-
sis. Robinson concludes by highlighting the potential benefits of employing
a socio-cognitive method: by mapping individual conceptualizations of a
polysemous category onto a variationist context, a dynamic picture of se-
mantic change in progress emerges.

In the final chapter in this section, "Applying word space models to sociolinguistics. Religion names before and after 9/11", **Yves Peirsman**, **Kris Heylen** and **Dirk Geeraerts** introduce word space models into cognitive semantics. They emphasize the importance of usage-based studies of lexical semantics based on more advanced techniques than just the extraction of examples from corpora. As the authors explain, by keeping track of the contexts in which a word appears, vector space models of lexical semantics approximate word meaning by modeling word use. To illustrate the usefulness of this computational-linguistic approach to lexical semantics, Peirsman, Heylen and Geeraerts present a case study based on a Dutch corpus of 300 million words and implement two types of word space models: a *document-based* and a *syntax-based* approach. The case study in question investigates how the use of religion names has changed after the attacks of 11 September 2001. The authors conclude that both the document-based and the syntax-based model show that *islam* has become distributionally more similar to words related to terrorism and politics and that *christendom*, by contrast, remains characterized by cultural and more positive dimensions.

2.2. Part two. Constructional variation

This section comprises three chapters that explore a variety of topics related to lectal variation in grammar and constructions.

The section opens with a contribution by **Benedikt Szmrecsanyi**, entitled "The English genitive alternation in a cognitive sociolinguistics perspective". In this study, which focuses on alternative constructions of genitive relations in British and American English, the author first selects the range of conditioning factors regarding choice of *of*-genitive vs. *s*-genitive whose univariate impact has been amply documented in linguistics. As a next step, the factors are subjected to a multivariate logistic regression model together with data extracted from three corpora of British and American English. In the analysis special attention is paid to how the external, sociolinguistic factors shape and determine the factor weights of the factors which are internal to language. In addition to the logistic regression analysis, the study furthermore relies heavily on visualization techniques such as cluster analysis and multidimensional scaling. Szmrecsanyi concludes that the most important language-external factor working on the English genitive alternation is the written/spoken text-type distinction and that the real-

time drift of written genitive choice systems, depending on their exact genre and on whether they are British or American, is differentially impacted by cultural phenomena such as colloquialization, Americanization and economization.

The next paper discusses the evolution of gender systems in the two national varieties of Dutch, Netherlandic Dutch and Belgian Dutch. In "(Not) acquiring grammatical gender in two varieties of Dutch", **Gunther De Vogelaer** uses a questionnaire to elicit instances of pronominal gender in 7-8 year-old children from a Netherlandic and a Belgian province. The results show that grammatical gender plays a much more important role in the Flemish children than in the Netherlandic data: While East Flemish children show attestations of feminine gender for non-animate count nouns and mass nouns, the children from the Dutch province only use feminine pronouns to refer to female humans or animates. On the basis of the qualitative data obtained, De Vogelaer deduces that three gender systems are operating in East Flemish children (the traditional three-gender system, the innovative dyadic grammatical system, and semantic gender) but that both the northern and southern systems are acquired as predominantly semantic systems. The author draws the tentative conclusion that northern and southern Dutch pronominal gender will ultimately converge in a system of semantic agreement.

The last chapter in this section likewise examines structural differences in national varieties of Dutch, but from different methodological and analytical perspectives. In "Lectal variation in constructional semantics: "Benefactive" ditransitives in Dutch", **Timothy Colleman** addresses the issue of lectal variation in constructional semantics through an exploration of semantic constraints on the benefactive ditransitive construction. Constructions form prototype categories exhibiting a cline from good to bad examples (*Could you pour me a cup of coffee* vs. *Could you taste me this wine*). Furthermore, constructions vary from language to language, either in nature or in the degree of productivity. As Colleman points out, in present-day standard Netherlandic Dutch, the benefactive ditransitive is a marked construction which is only possible with a handful of rather infrequent verbs related to food provision or preparation. However, the construction is productive in southern and eastern local dialect varieties with verbs from the semantic classes of creation and obtainment. In this chapter, the author investigates whether the wider semantic possibilities of the benefactive ditransitive in southern dialects manifest themselves in the standard language of Belgian speakers of Dutch as well. In order to test the distribution

of benefactive ditransitives in Netherlandic and Belgian Dutch, six frequent verbs of creation and obtainment were selected and searches were made for benefactive constructions in three different corpora representing various modes and registers of standard Netherlandic and Belgian Dutch. The results show that standard Belgian Dutch is more tolerant of the construction than standard Netherlandic Dutch. Colleman concludes that for the ditransitive to be possible in standard Netherlandic Dutch the preparatory act and the actual transfer must be contiguous, if not simultaneous, subevents. The semantic properties of abstract argument structure constructions can thus be subject to language-internal variation just like the semantic properties of lexical items.

2.3. Part three. Variation of lectal awareness and attitudes

The third and last part of the volume implies a thematic shift: topic-wisely we now turn to issues that pertain to lectal variation in relation to categorization, perception, awareness, attitudes, identities and acquisition. Within this group of three papers, we may note a difference with regard to the taxonomical level at which linguistic variation is studied. Whereas the last contribution deals with sound change in relation to attitudes within an apparently uniform lectal community, in the first two chapters focus is on the perception of lects perceived as whole units: language varieties in the form of (prototype) categories with their corresponding social and linguistic images, or social and linguistic stereotypes. In more specific terms, this cluster of contributions examines the acquisition of lectal awareness, differing attitudes towards lectal varieties and differing attitudes towards a specific lectal variable.

In "Lectal acquisition and linguistic stereotype formation", **Gitte Kristiansen** presents the design and results of a set of experimental studies carried out in order to examine lectal acquisition in young children. The aims of the investigation were to determine the crucial stages at which young children acquire receptive competence of lectal varieties at different levels of specificity and discuss potential predictors of the success rate: when do children acquire competence of lectal variation, at which levels of abstraction, and if such knowledge is experientially grounded, where does the knowledge stem from? The first experiment assessed the degree of identifiability of L1 accents in 150 Spanish children across three age groups. The second experiment examined the degree of correct identification of L2 ac-

cents. The results speak of consistent type-token relationship and a statistically significant increase in correct identification. As shown by data collected in an additional questionnaire, a high degree of success correlated with accents presenting a high degree of social stereotyping. In the light of the findings the author discusses the experiential grounding of linguistic stereotype formation and concludes that formal characteristics (such as phonetic salience) have fewer effects on correct identification than relative social salience (such as social stereotyping).

The starting-point of the next chapter, which examines dialect evaluation rather than dialect identification, is the opposite perspective. In "Investigations into the folk's mental models of linguistic varieties", **Raphael Berthele** questions the general view that negative or positive attitudes towards certain lects are due to cultural stereotypes and imposed norms and not to inherent characteristics of the varieties. In his study, Berthele employs visual stimulus mapping tasks and attribution tasks to elicit gestalt-based mental models of language varieties in the form of consistent patterns between visual traits and phonological features. Findings from these experiments (which investigate folk perception of Swiss German dialects) and from additional interviews provide evidence for a link between dialects with a high percentage of high vowels and chiseled, sharp and pointy forms. Berthele concludes that some of the evidence is consistent with the controversial "inherent value hypothesis", i.e. the claim that some language attitudes are due to inherent features of the systems and not only due to culturally or socially imposed norms, stereotypes or socio-cultural connotations. However, as the author points out, a focus on inherent and potentially universal features does not automatically entail a universalist, non-relativist position. Rather, the goal is to show how potentially universal perceptual mappings of sounds and forms interact with cultural and other mental models of social or ethnic groups, languages and varieties.

The volume closes off with the contribution "A cognitive approach to quantitative sociolinguistic variation: Evidence from *th*-fronting in Central Scotland", by **Lynn Clark** and **Graeme Trousdale**. The aim of this investigation was to examine the cognitive and social factors involved in a phonological change in progress (*th*-fronting) in East-Central Scotland. The data were collected over a period of 30 months from a group of 54 speakers who play in pipe bands in West Fife, Scotland and who form friendship groups who either favor or disfavor the use of the labiodental fricative under scrutiny. Data-gathering was thus carried out by means of ethnographic long-term participant observation, a technique commonly used in variation-

ist sociolinguistics. In order to reach a better understanding of the patterning of (th) the authors proceeded with a varbrul multivariate analysis into which linguistic, social and cognitive factor groups were included. In this chapter Clark and Trousdale resort to notions such as multiple inheritance, schematicity and full and partial sanction to explain how the linguistic variable in question can have a range of different social meanings within the same community.

It will be clear from these summaries that the contributions we have brought together in this volume are excellent illustrations of the two domains we identified above. The chapters in the first two sections are primarily concerned with *variation of meaning* (both lexical meaning and constructional meaning) and the third section is devoted to the *meaning of variation*, i.e. to the cognitive reality of meaningful variation in the minds of the language users. While the studies combined in this volume far from exhaust the domain, they do illustrate the potentialities of Cognitive Sociolinguistics: by joining forces in search of methodological refinement and descriptive expansion, Cognitive Linguistics and sociolinguistics may converge on a common goal – a detailed examination of variationist phenomena with the inclusion of semantic and cognitive factors.

References

Bybee, Joan, and Paul Hopper (eds.)
   2001   *Frequency and the Emergence of Linguistic Structure*. Amsterdam: John Benjamins.
Bybee, Joan
   2007   Diachronic linguistics. In *The Oxford Handbook of Cognitive Linguistics,* Dirk Geeraerts and Hubert Cuyckens (eds.), 945-987. New York: Oxford University Press.
Croft, William
   2009   Towards a social cognitive linguistics. In *New Directions in Cognitive Linguistics*, Vyvyan Evans and Stéphanie Pourcel (eds.), 395-420. Amsterdam: Benjamins.
Geeraerts, Dirk
   2005   Lectal variation and empirical data in Cognitive Linguistics. In *Cognitive Linguistics: Internal Dynamics and Interdisciplinary Interac-*

tion, Francisco J. Ruiz de Mendoza, and Sandra Peña Cervel (eds.), 163-189. Berlin/New York: Mouton de Gruyter

2006   A rough guide to Cognitive Linguistics. In *Cognitive Linguistics: Basic Readings,* Dirk Geeraerts (ed.), 1-28. Berlin/New York: Mouton de Gruyter.

Geeraerts, Dirk, and Stefan Grondelaers

1995   Looking back at anger: Cultural traditions and looking back at anger: Cultural traditions and metaphorical patterns. In *Language and the Cognitive Construal of the World*, John R. Taylor and Robert E. MacLaury (eds.), 153-179. Berlin/New York: Mouton de Gruyter.

Harder, Peter

2003   The status of linguistic facts: Rethinking the relation between cognition, social institution and utterance from a functional point of view. *Mind and Language* 18: 52-76.

Holland, Dorothy, and Naomi Quinn (eds.)

1987   *Cultural Models in Language and Thought.* Cambridge: Cambridge University Press.

Itkonen, Esa

2003   *What is Language? A Study in the Philosophy of Linguistics*. Turku: Åbo Akademis tryckeri.

Kemmer, Suzanne, and Michael Barlow

2000   Introduction. In *Usage-Based Models of Language*, Michael Barlow and Suzanne Kemmer (eds.), vii–xxviii. Stanford: CSLI Publications.

Kövecses, Zoltan

1995   Anger: Its language, conceptualization, and physiology in the light of cross-cultural evidence. In *Language and the Cognitive Construal of the World,* John R. Taylor and Robert E. MacLaury (eds.), 181-196. Berlin/New York: Mouton de Gruyter.

2005   *Metaphor in Culture: Universality and Variation.* Cambridge; New York: Cambridge University Press.

Kristiansen, Gitte

2003   How to do things with allophones: Linguistic stereotypes as cognitive reference points in social cognition. In *Cognitive Models in Language and Thought*, René Dirven, Roslyn M. Frank, and Martin Pütz (eds.), 69-120. Berlin/New York: Mouton de Gruyter.

Kristiansen, Gitte, and René Dirven (eds.)

2008   *Cognitive Sociolinguistics*. *Language Variation, Cultural Models, Social Systems.* Berlin/New York: Mouton de Gruyter.

Labov, William

1972   *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Langacker, Ronald W.
  1999   A dynamic usage-based model. In *Grammar and Conceptualization*,
            Ronald W. Langacker, 91-145. Berlin/New York: Mouton de Gruy-
            ter.
Lavandera, Beatriz
  1978   Where does the sociolinguistic variable stop? *Language in Society* 7:
            171-183.
Sinha, Chris
  2007   Cognitive linguistics, psychology and cognitive science. In *The
            Oxford Handbook of Cognitive Linguistics,* Dirk Geeraerts and Hu-
            bert Cuyckens (eds.), 1266-1294. New York: Oxford University
            Press.
Tomasello, Michael
  2003   *Constructing a Language: A Usage-Based Theory of Language
            Acquisition*. Cambridge, Mass.: Harvard University Press.
Tummers, José, Kris Heylen, and Dirk Geeraerts
  2005   Usage-based approaches in Cognitive Linguistics: A technical state
            of the art. *Corpus Linguistics and Linguistic Theory* 1: 225-261.
Verhagen, Arie and Van de Weijer, Jeroen
  2003   Introduction. In *Usage-Based Approaches to Dutch*, Arie Verhagen
            and Jeroen van de Weijer (eds.), 1–6. Utrecht: LOT.
Verhagen, Arie
  2005   *Constructions of Intersubjectivity: Discourse, Syntax, and Cognition*.
            Oxford: Oxford University Press.
Zlatev, Jordan
  2005   What's in a schema? Bodily Mimesis and the grounding of language.
            In *From Perception to Meaning: Image Schemas in Cognitive Lin-
            guistics,* Beate Hampe (ed.), 313-342. Berlin/New York: Mouton de
            Gruyter.

Part one.
Lexical and lexical-semantic variation

Heterodox concept features and onomasiological heterogeneity in dialects

Dirk Geeraerts and Dirk Speelman

Abstract

We examine the role of concept characteristics in the study of lexical variation among dialects: using a quantitative methodology, we show that onomasiological heterogeneity in a dialect area may be determined by the prototype-theoretical features of the concepts involved. More specifically, a regression analysis of data taken from a large lexical database of Limburgish dialects in Belgium and The Netherlands is conducted to illustrate that concept characteristics such as concept salience, concept vagueness and negative affect contribute to the lexical heterogeneity in the dialect data.

1.   Setting the question

To what extent do concept features determine the onomasiological heterogeneity that occurs in dialectological data? Onomasiological heterogeneity, in the way in which we intend it here, is the occurrence of synonymy across language varieties - in our case, across dialects. The phenomenon is common enough: neighbouring dialects, even closely related ones, need not feature the same word for the same concept. In dialectological and sociolinguistic research, the phenomenon is sometimes referred to as 'heteronymy' (Goossens 1969, Schippan 1992), and is then analyzed within the basic framework of variational language studies: lexical variation in a geographical or social continuum occurs because societal and material factors trigger the emergence and the subsequent interaction of different language systems. But from the point of view of Cognitive Linguistics, lexical variation may be determined by other factors next to such lectal ones: the

features of the concept itself may constitute a significant factor in the occurrence of lexical heterogeneity.

In particular, the linguistic development of prototype theory (see Geeraerts 2006) has brought to the attention a number of non-traditional semasiological features that may well influence the type of onomasiological variation that occurs across lectal boundaries. One such heterodox aspect is *vagueness* of meaning: if lexical meanings cannot always be easily distinguished, couldn't we perhaps expect more onomasiological heterogeneity when concepts are less clearly distinguishable? If the conceptual boundaries are unclear, different lexical choices between dialects may well occur more readily. Another such heterodox feature is conceptual *salience*: if concepts are psychologically more entrenched, couldn't we perhaps expect less onomasiological heterogeneity? If concepts are better known, highly familiar, more habitual, uniformity across dialects may be more easily achieved.

So, can we establish whether conceptual salience and conceptual vagueness significantly influence the occurrence of onomasiological heterogeneity? We will study the question by means of a statistical analysis of a large-scale database with dialectological data for the Limburgish dialects of Dutch. We will suggest an operational definition of vagueness, salience, and heterogeneity, and perform a multiple linear regression analysis on the data to test the hypothesis that certain concept features do indeed enhance onomasiological variation. In order to study the influence of concept features from as broad a perspective as possible, we will add one more concept feature to the analysis, viz. negative affect. Although we consider this to be an exploratory study, the overall results will be indubitable: concept features have a marked effect on heteronymy.

In disciplinary terms, the present study intends to contribute to three strands of research. First, it continues the line of quantitative, usage-based studies of onomasiological variation that was developed in our Leuven research group in the wake of Geeraerts, Grondelaers and Bakema (1994) and Geeraerts (1997): see Geeraerts, Grondelaers and Speelman (1999), and compare Speelman, Grondelaers and Geeraerts (2003) for an introduction to the methodological background.

Second, it adds a topic for investigation to the field of quantitative dialectology, as it is illustrated by researchers like Goebl (2006), Kretzschmar (2006), Nerbonne and Kleiweg (2007). Within this approach, semantic factors have as yet hardly been taken into consideration. With the present study, we would like to show how relevant it can be to do so.

And third, the paper is a contribution to dialectological studies within Cognitive Linguistics. Because language-internal variation has only recently come to the attention of Cognitive Linguistics (see Kristiansen 2003, Geeraerts 2005, Bernárdez 2005, Kövecses 2005, Kristiansen and Dirven 2008), it comes as no surprise that non-standard variants of a language have for long decades been neglected in cognitive linguistic studies. In spite of pioneering but largely isolated efforts such as Moerdijk and Geeraerts (1991), Swanenberg (2000), Nilsson (2001), Berthele (2002, 2004, 2006), Sharifian (2005), Szelid and Geeraerts (2008), there is no standing tradition of dialectological research in Cognitive Linguistics. That is to be regretted, because the inspiration could well be mutual. On the one hand, dialectological data raise the question whether differences of culture and conceptualization, one of the theoretical centers of attraction of Cognitive Linguistics, could be detected language-internally (and not just, according to the usual perspective of Cognitive Linguistics, across languages). On the other hand, the usage-based nature of Cognitive Linguistics challenges the traditional methodological focus of dialectology on language structure rather than language use. In this paper, we offer one more example of what such a cognitive linguistic dialectology could look like.

2.   Sketching the design

How then shall we try to answer our basic research question? In order to specify the design of our study, we need to say something about the material that we will be using as the descriptive basis of our investigation, about the explanatory variables that we will include, and about our operationalization of the response variable. In the presentation of the design and the results, we will keep the exposé deliberately brief and fairly abstract, i.e. in order to concentrate on the essential architecture of the approach, we will not attempt to illustrate each successive step with numerous examples, nor with technical details about the calculations used. For additional detail, we refer to Speelman and Geeraerts (2009), a follow-up study that compares the methodology presented here with alternative approaches to lexical dialectometry.

2.1. The material basis

We will investigate the instalment devoted to the human body of the *Woordenboek van de Limburgse Dialecten*, or WLD. The WLD is a large-scale dictionary devoted to the Limburgish dialects as spoken in the Dutch province of Limburg and the Belgian province of Limburg. The initiative for the WLD was taken by Anton Weijnen of the university of Nijmegen in the 1960s. The first installment was published in 1983, and the dictionary is currently being completed at the universities of Nijmegen and Leuven. (The Leuven collaborator on the project, Ronny Keulen, has been an indispensable help to us in the preparation of this study, by making available the electronic database behind the dictionary and by guiding us through its design.) The dictionary is organized thematically. In topical instalments like 'The household: Food and drink' or 'Agrarian terminology: Grassland farming', the traditional terminology of the dialects is described and charted - in a fairly literal sense, to the extent that the geographical distribution of the various terms is indicated by means of dialectological maps. The instalment we are using is WLD III.1.1. 'The human being as an individual: The human body' (Keulen 2004). At present, the WLD is made available online (http://www.ru.nl/dialect/wld/).

The materials included in the WLD come from many sources. The bulk of the data was collected by means of questionnaires specifically designed for the dictionary project, but these questionnaires were complemented with material from older dialect surveys and sundry sources. However, because we would like to base our study on maximally homogeneous material in terms of age and geographical scope, we will only be using the questionnaires N10, N106, N107, N108 and N109. All of these were applied fairly systematically over the entire geographical region covered by the dictionary, in a fairly recent period (roughly, the last third of the previous century). This restriction means, for instance, that data culled from research monographs devoted to a single local dialect, or data from larger scale dialect surveys undertaken in the 1930s are not included in our analysis.

We need to insist on the systematicity with which the data are collected, because we would otherwise enhance the ambiguity of non-responses. If we do not find any name at all for a given concept in a given place, that could in principle have two reasons: either the question was never asked, or no term was known to the informants (because the concept was unfamiliar, or because there was a lexical gap in the dialect). The latter situation is one

that would be of interest to us (could it be the case that lesser known concepts are lexically more heterogeneous than others?), but we will only be able to investigate such effects if we can rule out, or at least minimize, the occurrence of the former situation.With the restrictions that we applied, we can base our investigation on a database of 32591 tokens of lexical items, divided over 206 concepts and 201 geographical places.

2.2. The explanatory variables

The concept features that we will incorporate into the analysis are threefold: the salience of the concepts, the vagueness of the concepts, and the negative affect of the concepts. Vagueness and salience are features that may be typically associated with prototype theory and a cognitive linguistic conception of lexical meaning. Negative affect, on the other hand, is a more traditional semantic feature. Vagueness and negative affect will be included in the form of a single operationalization, but salience will be measured in the form of three distinct factors (which will be considered separately in the analysis): the habituality of the concept, the number of null responses for the concept, and the number of multiword expressions featuring among the designations of the concept. Let us now look at each of these five factors (the three salience factors, vagueness, and negative affect) in turn, and indicate why exactly we have included them and how we have tried to measure them.

1. The *lack of familiarity* of a concept is relevant because we suspect that less habitual concepts increase lexical uncertainty among language users, and decrease the probability of uniformity across dialects. If a concept is less common, it is communicatively less prominent, and the possibility (or perhaps also the necessity) for standardization is more restricted. Our operational measurement of lack of conceptual habituality is relatively basic: we have conducted a survey among seven members of our research group, asking them to rate the 206 bodily concepts in the database on a five point scale of habituality, with 1 indicating no risk of unfamiliarity, and 5 a high risk of unfamiliarity. The results we get on the survey are consistent and intuitively plausible. At the lower end of the scale, for instance, we find such concepts as KNOKKELKUILTJES 'the little dents between the knuckles of the hand', BLOEDWEI 'blood plasma, the liquid component of blood', LEVEND VLEES ONDER DE HUID 'living flesh underneath the skin', VOOR-

VOET 'front part of the foot', AFHANGEND KUIFJE (BIJ KORTGEKNIPT HAAR) 'forelock, cowlick (in case of short hair)' in contrast with habitual ones like KEEL 'throat', KNIE 'knee', MIDDELVINGER 'middle finger'.

2. The inclusion of the *number of observational gaps* is motivated by the idea that a high number of places without responses may be an indirect indication of lack of familiarity with the concept: if the informants are unable to provide a name for a concept, then that may mean that they simply do not know the concept (and that, in turn, may be a significant factor contributing to overall onomasiological heterogeneity). However, this reasoning assumes that the concept was indeed included in the survey. By restricting the database to the questionnaires N10, N106, N107, N108 and N109, we have tried to ensure that this is indeed the case, but we cannot be entirely certain about the observational systematicity of the materials. An observational gap, in other words, is ambiguous between actual unfamiliarity and an inconsistency in the survey procedure. Some caution with regard to this factor will be in order, then, all the more so since there might be a mathematical effect on heterogeneity in the opposite direction of what we expect (fewer measurements may lead to a smaller number of names, thus possibly reducing heterogeneity). In operational terms, the number of observational gaps is quantified straightforwardly as the absolute number of places (out of the total of 201 places) in which no names were given for the concept at hand. Concepts with few responses are, for instance, SLECHT GROEIEN 'not to grow well, to grow slowly', GELUIDLOZE WIND 'noiseless fart', KAAKGESTEL 'jawbone' and HUIG 'uvula'.

3. The *number of multiword expressions* in the onomasiological range of a concept may be considered an indication of lack of salience for two reasons. (Throughout the paper, we use the concept 'onomasiological range' to refer to the total set of expressions that occur as designations of the concept. If we take into account the relative frequency of those expressions within the onomasiological range, we talk about an 'onomasiological profile'.) First, the basic level hypothesis (Berlin and Kay 1969, Berlin 1978) suggests that cross-linguistically basic concepts are typically referred to with short words. Second, multiword answers may derive from the fact that people answer with a periphrastic description of the concept either because there is no name for the concept in their dialect or because they don't know it: informants may creatively invent names on the spot because they don't know what to answer. According to the first reason, multiword expressions

constitute a structural reflection of lower concept salience; according to the second reason, multiword expressions reflect the fact that the concept is not structurally entrenched in the lexicon at all. Examples of multiword expressions in the database include *haar wie een stekelvarken* 'hair like a porcupine', *haar wie stro,* 'hair like straw', *steil haar* 'flat, straight hair', *stijf haar* 'stiff hair', for the concept BORSTELIG HAAR 'brushy hair'. As a measurement, we take the proportion (at the token level) of the number of multiword answers in the total set of answers for a concept.

4. The vagueness of concepts is measured in terms of their *lexical non-uniqueness*: if a term for naming the concept at hand can also be used for naming another concept, then this may be seen as an indication that the concept at hand is not always easily distinguishable from the other concept. We quantify lexical uniqueness of a concept as the number of lexical types in the onomasiological range of that concept that also features in the set of expressions associated with a different concept. As an example, consider the following overview of the expressions found for the concept LIES 'groin'. The first column lists the lexical expressions we find, the second the frequency with which they are found, and the third mentions whether the same lexical expression may also be found as expressing another concept. In this example, then, five out of nine types are non-unique.

Table 1. Expressions found for the concept LIES 'groin'

lies	178	*lies* also used for BEKKENHOLTE 'pelvic cavity'
de dun	17	
vlim	6	*vlim* also used for WIMPER 'eyelash'
lende	2	*lende* also used for LENDE 'loin'
liest	2	
hees	1	*hees* also used for KNIEHOLTE 'knee pit'
lee	1	*lee* also used for LENDE 'loin'
liesje	1	
var	1	

The example makes clear that alternative operationalizations of lexical non-uniqueness will have to be envisaged in future research. Next to the introduction of a proportional token-based rather than type-based measurement, attention will have to be devoted to the distinction between vagueness and

polysemy. Some of the overlaps in the example, as in the case of *vlim*, do not constitute prime examples of the phenomenon that motivates the introduction of vagueness in the analysis. We are interested in the effect of vagueness because we assume that conceptual unclarity, like the fuzziness of the borderline between one concept and the other, may lead to lexical heterogeneity. While this may apply to the relationship between BEKKEN-HOLTE 'pelvic cavity'and LIES 'groin', which are plausibly easy to confuse, a conceptual confusion between 'eyelash' and 'loin' seems unlikely. If a workable criterion for singling out such cases can be found (but see Geeraerts 1993), we will be able to investigate whether conceptual overlaps of the *vlim* type have a different effect from those of the *lies* type. For the present exploratory purposes, however, we restrict the analysis to the course-grained measure described above.

5. The inclusion of *negative affect* is motivated by the recognition that taboo leads to rich synonymy (Allan and Burridge 1988). Given domains like procreation and defecation, taboo is obviously relevant for the lexical field of the human body. Negative affect is not restricted to such obvious taboo areas, however. The dictionary contains many questions in which it is explicitly asked to give pejorative terms for a certain concept, i.e. we find concepts that are explicitly marked as negative and that were surveyed as such. Examples are NEUS (SPOTNAAM) 'nose (pejorative)', GEZICHT (SPOT-NAAM) 'face (pejorative)', and HOOFD (SPOTNAAM) 'head (pejorative)'. In practical terms, we did not start from these labels, but we used the same method for the identification of negative affect as for the identification of lack of familiarity: seven members of our research group rated the 206 bodily concepts in the database on a five point scale of negative affect, with 1 indicating no negative affect, and 5 a strong negative affect. The results are again consistent and plausible: examples of concepts with strong negative affect are AARSSPLEET 'anal cleft', GELUIDLOZE WIND 'noiseless fart', KWIJL 'drool', or PAPPERIG PERSOON 'fat, plump person'.

2.3. The response variable

Lexical heterogeneity, the dependent variable in the investigation, is defined as a complex factor. If, in fact, we consider which phenomena can point to heterogeneity, we will not only wish to take into account lexical diversity (the existence of different words for naming a concept) but also

geographic fragmentation: there is more heterogeneity when there is non-homogeneity (or 'scatter') in the geographic distribution of those different words. Heterogeneity, then, may be defined as the product of lexical diversity and geographic fragmentation. Geographic fragmentation, however, needs to be analyzed further. On the one hand, it is linked to the geographical range of the terms: a smaller average range indicates higher fragmentation. On the other hand, it is linked to the dispersion of the terms within that range: holes in the distribution indicate higher fragmentation. Let us now have a closer look at each of the three relevant phenomena (diversity, range, dispersion) and their operationalization.

1. Lexical *diversity* is straightforwardly defined as the number of different types (including multiword expressions) in the onomasiological range of a concept. In follow-up research, we intend to explore measures of diversity in which the distribution of tokens over types is taken into account as well (like a type/token ratio, or the measure of internal uniformity used in Geeraerts, Grondelaers and Speelman 1999).
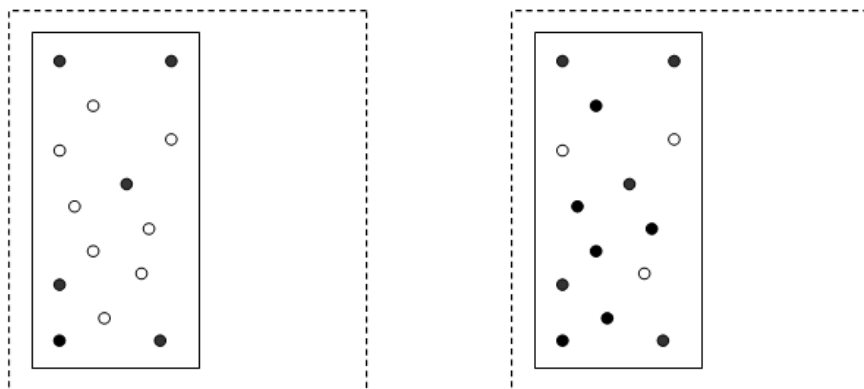
Figure 1. A schematic representation of geographical dispersion

2. The *dispersion* of a concept is illustrated by means of Figure 1: dispersion is the extent to which the lexical attestations of a concept cover the total set of observational points under consideration. If we think of the dotted rectangle in the figure as the total area under consideration (in our case, the Limburgish dialect area), then the solid line may be used to indicate that part of the global area in which a given concept appears. (In most cases in our database, the concepts appear in the Limburgish region in its entirety.)

We can then think of each of the dots as places - towns and villages - in that area. A black dot indicates a place where we find an attestation for a given concept, and a white dot indicates a place where we get a null observation for the concept, i.e. where the concept is represented by no lexical item. The situation on the left hand side of the figure is intuitively more dispersed than the situation on the right, but how can we turn that intuition into a quantitative measurement? We express dispersion as a proportion between average distances.

First, we take the distance from one observation of a term to the immediately neighboring observation, i.e. the closest other observation of that term. The distances are geographical distances, based on the latitude and altitude of the places. We do this for all other observations of the term and calculate the average distance to the immediately neighboring term observation. Informally, this is the average distance between a black dot and another black dot representing the same lexical item.

Second, we take the distance from one observation of a term to the immediately neighboring observation of the concept, i.e. the closest other observation of that concept, regardless of whether the nearest neighbor appears with the same lexical item or not. We do this for all other observations of the term and calculate the average distance to the immediately neighboring concept observation.

Third, for each term, we take the proportion of the two averages that we just described. This measure yields the dispersion for a single term in the onomasiological range of a concept, but we are obviously interested in the overall dispersion for the concept. That is why we calculate a weighed average of the measures of dispersion of the individual terms: we average over the dispersion of the terms, but we use a weight factor that corresponds to the relative frequency of the term in the onomasiological profile of the concept.

3. The *range* of a concept is illustrated by Figure 2. Each of the solid boxes indicates the maximal geographical range of a given term for a given concept, regardless of the dispersion within that maximal area of occurrence. The situation on the left hand side of the figure is more fragmented than the situation to the right: the average area covered by the various terms is smaller to the left than the average area covered to the right. In practical terms, range is calculated in the following steps. First, we calculate the surface that is covered by the attestations of one term. Second, we calculate the proportion of that surface to the surface that is covered by the concept

as a whole. Third, we calculate this proportion for each term in a concept and then take the weighted average proportion. As before, we are weighting terms by their token size within the onomasiological profile of the concept.
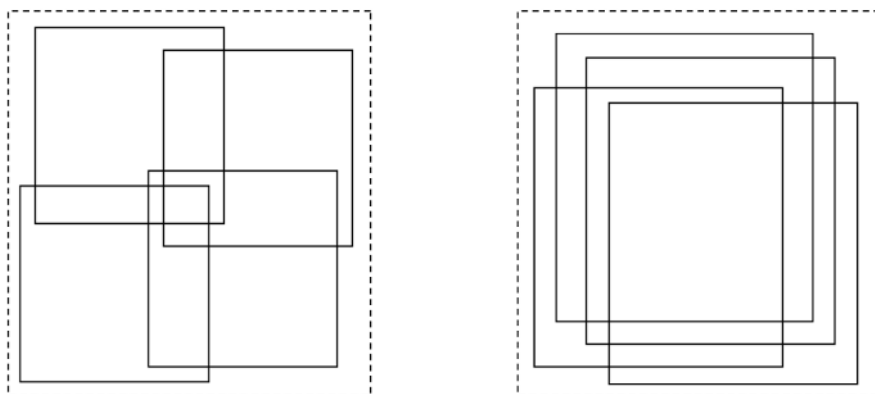
Figure 2. A schematic representation of geographical range

Once we have a measure for diversity, dispersion, and range, we can go back to our earlier definition of heterogeneity as the product of diversity and geographic fragmentation. Because geographic fragmentation rises as dispersion increases, and diminishes as range increases, fragmentation may now be defined as the proportion of dispersion and range. The overall formula for heterogeneity then takes the following form:

heterogeneity = diversity x (dispersion / range)

3. Analyzing the data

The response variable and the five explanatory variables are subjected to a multiple linear regression analysis. The results of the analysis are presented in Table 2. The abbreviations used for the predictor variables as follows.

MISSING.PLACES: number of observational gaps
LACK.FAMIL: (assumed risk for) lack of familiarity
PROP.MULTIWORD: proportion of multiword answers
NON.UNIQUENESS: occurrences of names in other concepts
NEG.AFFECT: (assumed) negative affect

Table 2. The results of the multiple linear regression analysis

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.061618   0.350465   3.029  0.00281 **
missing.places  -0.005888   0.001984  -2.968  0.00341 **
lack.famil       0.740298   0.142952   5.179 5.94e-07 ***
prop.multiword   2.782169   0.428651   6.491 8.04e-10 ***
non.uniqueness   0.053341   0.007283   7.324 7.78e-12 ***
neg.affect       0.540066   0.120095   4.497 1.23e-05 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.216 on 180 degrees of freedom
Multiple R-squared: 0.6232
Adjusted R-squared: 0.6128
F-statistic: 59.55 on 5 and 180 DF,  p-value: < 2.2e-16
```

Before we have a closer look at the results, a number of technical remarks need to be made; these will be relevant only for those readers who are familiar with the technical apparatus of a regression analysis. First, because the residual values are not normally distributed when heterogeneity as such is used as the response variable, the regression analysis is based on the logarithm of heterogeneity. Second, to avoid cases of extreme data sparseness, we have restricted the analysis to concepts that are attested in at least ten places. This leaves us with 186 of the original 206 concepts. Third, two interactions need to be mentioned in addition to the basic results. For one thing, lack of familiarity enhances heterogeneity only in the case of low or medium non-uniqueness, but it has no effect in the case of extremely high non-uniqueness. The second interaction is similar: negative affect triggers heterogeneity only in the case of low or medium non-uniqueness, but it has no effect in the case of extremely high non-uniqueness. Because both interactions do not substantially influence the analysis (neither from a technical nor from an interpretative point of view), we consider it legitimate to simply focus on the model without interactions in the rest of the discussion - even though the model with the interactions is intrinsically more accurate. Fourth, we find 3 outliers and 19 influential observations in the data set. Leaving these 22 observations out of the analysis yields a slightly better model than the one presented in the table: we reach an adjusted R-squared of 0.7173, and the standard error for residuals decreases slightly. However,

because these differences are not fundamental, neither from a technical or an interpretative point of view, we again concentrate on the original model that includes the 23 observations.

Now, what can we conclude from the results as presented in the table? In the first place, with more than 61% of explained variation (see the figure indicated by Adjusted R squared), the model may be considered a good one. With a significance value of less than 0.00001 for the F-test, the model performs significantly better than an intercept only model, i.e. a model in which the effect of the explanatory factors is not taken into account. In the second place, all the factors that we included as explanatory variables appear to have a significant effect. (This is indicated by the significance factors in the final column of the table.) This is a crucial finding, because it corroborates our initial and fundamental assumption that concept features, and more specifically, heterodox concept features, influence lexical heterogeneity. In the third place, when we turn to the first column of figures, we observe that all factors have a positive effect on lexical heterogeneity, except for the factor 'missing places'. (This follows from the fact that all the estimates are positive numbers, except the estimate for 'missing places'.) This means that heterogeneity increases as a concept is less familiar, exhibits more multiword answers, overlaps more with other concepts, and has a higher negative affect, but that heterogeneity decreases as the number of places with zero observations rises. Except for the latter, these observations are entirely in accordance with the hypotheses that we put forward.

The different behavior of the number of observational gaps is not a total surprise, however. When we introduced the factor, we mentioned that observational gaps could be ambiguous, to the extent that they could either result from an unsystematic survey technique, or from lack of familiarity with the concept. A calculation of the effect of the factors (which we will not present in detail here) shows that the effect of the number of observational gaps is the weakest of all the factors considered, which we take as an extra indication that the factor needs to be scrutinized in more detail in the course of further investigations.

4.   Suggesting further prospects

The central conclusions to be drawn from our exploratory investigation into the sources of lexical heterogeneity in the *Woordenboek van de Limburgse Dialecten* are clear. Taking into account non-orthodox concept features

(salience and vagueness) helps to account for lexical heterogeneity in dialect databases: studying dialectological lexical heterogeneity purely from the point of view of geographic differentiation is too restrictive. And as the influence of negative affect shows, concept features need to be taken into account more generally: it is not just the features highlighted by prototype theory and cognitive semantics that turn out to be relevant for the explanation of heterogeneity.

On top of these theoretically relevant conclusions, there is an important methodological conclusion to be highlighted: more advanced forms of quantitative analysis, like in our case, multiple regression analysis, clearly help to cope with the complexity of dialectological lexical materials.

Given the apparent fruitfulness of the approach illustrated here, we may conclude with the identification of prospects for further research. Quite a number of perspectives open up. In the first place, we may try out alternative forms of the study as it was presented here. As we indicated earlier, alternative operationalizations of the factors should be explored, like a token-based rather than a type-based measure of lexical non-uniqueness, or similarly, a token-based measure of diversity. In the same vein, we may consider an alternative calculation of range and dispersion on the basis of 'number of places' instead of surfaces and distances, and we should consider measures to distinguish between vagueness and polysemy in the calculation of non-uniqueness. The design may be varied in still other respects: we may split up the results for different geographic regions (do the variables work in the same way in the Belgian province of Limburg as in the Dutch province of Limburg?), or we may have a separate look at the two components of heterogeneity, i.e. diversity and geographic fragmentation, instead of combining them in a single measure.

In the second place, we may extend the study beyond its present limits by taking into account other regions: if we take similar data from the dictionaries of the Brabantish dialects and the Flemish dialects (which are compiled in parallel to the dictionary of the Limburgish dialects), can we confirm our findings? And even more appropriately, given our interest in semantics, we may envisage an extension towards other lexical fields, as represented by other installments of the dictionary: do the various factors that we have identified for the lexical field of human body play the same role in other fields, and what is the role of the field itself?

References

Allan, Keith, and Kate Burridge
    1988   Euphemism, dysphemism, and cross-varietal synonymy. *LaTrobe working papers in linguistics* 1: 1-16.

Berlin, Brent
    1978   Ethnobiological classification. In *Cognition and Categorization*, Eleanor Rosch, and Barbara B. Lloyd (eds.), 9-26. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Berlin, Brent, and Kay, Paul
    1969   *Basic Color Terms: Their Universality and Evolution*. Berkeley/Los Angeles: University of California Press.

Bernárdez, Enrique
    2005   Social cognition: variation, language, and culture in a cognitive linguistic typology. In *Cognitive Linguistics: Internal Dynamics and Interdisciplinary Interaction*, Francisco J. Ruiz de Mendoza, and Sandra Peña Cervel (eds.), 191-222. Berlin/New York: Mouton de Gruyter.

Berthele, Raphael
    2002   Learning a second dialect: A model of idiolectal dissonance. *Multilingua* 21: 327-344.
    2004   The typology of motion and posture verbs: A variationist account. In *Dialectology Meets Typology. Dialect Grammar from a Cross-Linguistic Perspective*, Bernd Kortmann (ed.), 93-126. Berlin/New York: Mouton de Gruyter.
    2006   *Ort und Weg. Die sprachliche Raumreferenz in Varietäten des Deutschen, Rätoromanischen und Französischen*. Berlin/New York: Walter de Gruyter.

Geeraerts, Dirk
    1993   Vagueness's puzzles, polysemy's vagaries. *Cognitive Linguistics* 4: 223-272.
    1997   *Diachronic Prototype Semantics*. Oxford: The Clarendon Press.
    2005   Lectal variation and empirical data in Cognitive Linguistics. In *Cognitive Linguistics: Internal Dynamics and Interdisciplinary Interaction*, Francisco J. Ruiz de Mendoza, and Sandra Peña Cervel (eds.), 163-189. Berlin/New York: Mouton de Gruyter.
    2006   Prototype theory: Prospects and problems of prototype theory. In *Cognitive Linguistics: Basic Readings*, Dirk Geeraerts (ed.), 141-166. Berlin/New York: Mouton de Gruyter.

Geeraerts, Dirk, Stefan Grondelaers, and Peter Bakema
    1994   *The Structure of Lexical Variation. Meaning, Naming, and Context*. Berlin/New York: Mouton de Gruyter.

Geeraerts, Dirk, Stefan Grondelaers, and Dirk Speelman

    1999   *Convergentie en divergentie in de Nederlandse woordenschat. Een onderzoek naar kleding- en voetbaltermen.* Amsterdam: Meertens Instituut.

Goebl, Hans

    2006   Recent advances in Salzburg dialectometry. *Literary and Linguistic Computing* 21: 411-435.

Goossens, Jan

    1969   *Strukturelle Sprachgeographie.* Heidelberg: Carl Winter.

Keulen, Ronny

    2004   *Woordenboek van de Limburgse Dialecten*, Deel III, sectie 1, afl. 1, *De mens als individu, Het menselijk lichaam.* Utrecht: Gopher.

Kövecses, Zoltan

    2005   *Metaphor in Culture: Universality and Variation.* Cambridge/New York: Cambridge University Press.

Kretzschmar, William A.

    2006   Art and science in computational dialectology. *Literary and Linguistic Computing* 21: 399-410.

Kristiansen, Gitte

    2003   How to do things with allophones: Linguistic stereotypes as cognitive reference points in social cognition. In *Cognitive Models in Language and Thought*, CLR 24, René Dirven, Roslyn M. Frank, and Martin Pütz (eds.), 69-120. Berlin/New York: Mouton de Gruyter.

Kristiansen, Gitte, and René Dirven (eds.)

    2008   *Cognitive Sociolinguistics. Language Variation, Cultural Models, Social Systems.* CLR 39. Berlin/New York: Mouton de Gruyter.

Moerdijk, Alfons, and Dirk Geeraerts

    1991   Het systematische dialectwoordenboek en de lexicologische gebruiker. *Taal en Tongval* 43: 58-104.

Nerbonne, John, and Peter Kleiweg

    2007   Toward a dialectological yardstick. *Journal of Quantitative Linguistics* 14: 148-166.

Nilsson, Anneli

    2001   Semantic shift among adjectives in the Southern-Swedish dialect of Scania. *Travaux de l'Institut de Linguistique de Lund* 39: 227-240.

Schippan, Thea

    1992   *Lexikologie der deutschen Gegenwartssprache.* Tübingen: Niemeyer.

Sharifian, Farzad

    2005   Cultural conceptualisations in English words: A study of Aboriginal children in Perth. *Language and Education* 19: 74-88.

Speelman, Dirk, and Dirk Geeraerts
    2009   The role of concept characteristics in lexical dialectometry. *International Journal of Humanities and Arts Computing* 2: 221-242.

Speelman, Dirk, Stefan Grondelaers, and Dirk Geeraerts
    2003   Profile-based linguistic uniformity as a generic method for comparing language varieties. *Computers and the Humanities* 37: 317-337.

Swanenberg, Jos
    2000   Lexicale variatie cognitief-semantisch benaderd. Over het benoemen van vogels in Zuid-Nederlandse dialecten. Ph.D. diss., Katholieke Universiteit, Nijmegen.

Szelid, Veronika, and Dirk Geeraerts
    2008   Usage-based dialectology. Emotion concepts in the Southern Csango dialect. *Annual Review of Cognitive Linguistics* 6: 23-49.

Measuring and parameterizing lexical convergence and divergence between European and Brazilian Portuguese

Augusto Soares da Silva

Abstract

Following the model provided by the sociolectometrical and sociolexicological study that Geeraerts, Grondelaers, and Speelman (1999) performed for Netherlandic and Belgian Dutch, the present study investigates the relationship between the vocabulary of European Portuguese and that of Brazilian Portuguese. Focusing on the lexical field of clothing terms and that of football terms, two main issues: are addressed by means of the quantitative methodology developed by Geeraerts et al.: whether the two national varieties of Portuguese have gone through a process of lexical convergence or divergence in the last 60 years; and which internal linguistic parameters influence this process.

1. Introduction

This paper aims to present the main aspects and results of a research study into the lexical relationship between European Portuguese (EP) and Brazilian Portuguese (BP). There are two main issues: (i) to know whether these two national varieties of the Portuguese language have gone through a process of lexical convergence or divergence in the last 60 years; and (ii) to find out how internal linguistic parameters influence these global convergence or divergence trends.[1] Both issues involve a diachronic analysis, an external one for the former and an internal one for the latter. Internal parameters include endo-/exogenousness and foreign influence (loanwords). Other item-related and concept-related features will also be analyzed. Additionally, this study provides some insights into the synchronic issue of lexi-

cal stratification in both varieties so as to evaluate whether distance between the standard and substandard strata is greater in BP or EP.

Among the hypotheses about the relationships between EP and BP, it is conjectured (i) that there is an increasing influence of BP on EP (and African varieties) as a result of the popularity of Brazilian soap operas and football, as well as Brazilian emigration; (ii) that BP is more receptive to loanwords; (iii) that there is a greater distance between formal and informal registers in BP than in EP; and (iv) that despite the lack of clear hypotheses stated in the literature, a progressive and inevitable fragmentation of the Portuguese language is taking place (i.e. the hypothesis of divergence).

The present investigation is concerned with onomasiological variation involving denotational synonyms. The empirical background consists of several thousand observations of the use of alternative terms that designate 43 nominal concepts from the lexical fields of football and clothing. This corpus-based onomasiological investigation follows up on the original sociolectometrical and sociolexicological study that Geeraerts, Grondelaers and Speelman (1999) performed for Netherlandic and Belgian Dutch. Similar to Geeraerts et al. (1999), this research is based on a cognitive sociolinguistics perspective (Geeraerts 2005; Kristiansen and Dirven 2008), an emerging area in Cognitive Linguistics which is committed to the study of the relationship between cognition and language-internal diversity, in other words "lectal" variation.

This paper is divided into six sections. Following this introduction, the second section provides the essential elements of the theoretical and methodological framework of this cognitive and quantitative sociolexicological study of the Portuguese language. The third and fourth sections present uniformity results and item/concept-related results respectively. Special emphasis is placed on the mutual impact of uniformity and item-related features. Additionally, the fifth section gives some insights into the question of stratification. The conclusions and topics for further research are given in the last section. In this way, we hope to contribute to the development of Portuguese sociolinguistics and particularly to the issue of convergence and divergence between the European and Brazilian varieties of the language (see Soares da Silva 2005, 2008a).

2. Background and methodology

Differences between EP and BP exist at all levels of linguistic structure. Innovative and conservative trends have emerged in both varieties, such that tradition is not the privilege of EP nor is innovation the privilege of BP.

The issue of a "Brazilian language" is back on the agenda and has awakened those same passionate attitudes triggered in the past on three occasions, namely Brazil's independence in 1822, the period of Romanticism and the period of Modernism. One example is the famous essay written by Brazilian linguist Bagno (2001). However, linguistic purism is growing stronger in Brazil nowadays. Recently, a Federal bill made provision for forbidding the use of foreign words and stipulated the payment of fines for those who breached the law (see Faraco 2004).

BP presents a situation of diglossia – there is a clear distance between the idealized and prescriptive traditional norm and the real norm (or norms) used in big city centers – and also a wide dialectal continuum (Mattos e Silva 2004), while an increasing standardization of EP has been observed since the 1974 democratic revolution. BP is now facing two major challenges: a sociolinguistic dilemma (due to great regional and social variation) and a didactic dilemma (teaching the language to a soaring population). A population of 220 million Brazilians is foreseen in the next 15 years, that is to say, a 40 million increase in population (Castilho 2005).

As mentioned in the previous section, the four hypotheses about lexical relationships between EP and BP are: (i) increasing influence of BP on EP; (ii) stronger foreign influence in BP; (iii) greater stratification in BP; and (iv) divergence between EP and BP. As for the hypothesis of divergence, a well-known journalist wrote in a Portuguese reference newspaper:

> Continua a haver uma só língua Portuguesa, mas nos últimos 50 anos, digamos, têm-se acentuado as diferenças na sintaxe e no léxico – sobretudo no vocabulário corrente. ['There is still one Portuguese language only, but let's say that, in the last fifty years, the differences in syntax and lexicon have been emphasized – above all in current vocabulary'] (Belard 2001)

The anticipated fragmentation of the Portuguese language is clearly put forward by a distinguished historical linguist on the grounds of what happened to Latin in the declining Roman Empire:

> o modelo latino parece legitimar a previsão de que línguas como o inglês, ou o português, faladas em espaços muito vastos, habitados por povos que

nada têm a ver historicamente com o berço da língua, se acham destinadas a uma progressiva fragmentação interna. ['The Latin model seems to give legitimacy to the forecast that languages like English or Portuguese, which are spoken in extended areas inhabited by people who are not historically related to the cradle of the language, are destined to progressive internal fragmentation'] (Castro 1986: 45)

Nevertheless, even the most detailed comparative studies of EP and BP (Brandão and Mota 2003; Peres and Kato 2001, 2004) do not tackle the issue of convergence/divergence.

The object of study is a specific form of lexical variation, namely *formal onomasiological variation* (Geeraerts, Grondelaers, and Bakema 1994). Onomasiological variation is *formal* when many different terms are used to refer to the same entity. This variation is not due to a different conceptual classification of the same entity, but rather to the use of many different synonymous terms, i.e. denotational synonyms, to refer to the same concept. For example, the variation between *avançado*, *atacante* and *dianteiro* for the referent 'forward' is formal, while the variation between *avançado* 'forward' and *jogador* 'player' is conceptual. Formal onomasiological variation is particularly interesting from a sociolinguistic point of view because the use of denotational synonyms generally gives some hints as to the relationships existing between language varieties. Indeed, denotational synonyms are likely to reveal sociolinguistic differences, i.e. regional, social, stylistic and pragmatic-discoursal differences, and it is these differences that motivate the very existence and competition of *lectal* varieties. Formal onomasiological variation, of which contextual variation is an integral part, is essentially the most specific subject of *sociolexicology* and of this study.

Data was gathered from the lexical fields of football and fashion/clothing due to their popularity and the fact that they are susceptible to the influence of foreign languages. A third lexical field is still under study – health. The empirical background of this sociolexicological study consists of several thousand observations of the use of denotational synonyms which designate 43 nominal concepts from football and clothing terminologies.

Material was extracted from three different sources: (i) sports newspapers and fashion magazines from the early years of the 1950s, 1970s and 1990s/2000s; (ii) Internet Relay Chat (IRC) channels related to football; and (iii) labels and price tags pictured from shop windows in two Portuguese and Brazilian towns respectively. Material gathered from (i) can help to answer the question of convergence/divergence while material collected

from (ii) and (iii) shed light on the question of stratification. All the material from (i) and (iii) was manually extracted. Data referring to the Brazilian variety was collected from the two largest cities in the country, namely São Paulo and Rio de Janeiro.

The sub-corpus of football contains 2.7 million tokens selected from 8 newspapers (4 Portuguese and 4 Brazilian newspapers) and 15 million tokens collected from Internet chats. The sub-corpus of clothing extends to 1.2 million tokens gathered from 24 fashion magazines (14 Portuguese and 14 Brazilian magazines) and 1,300 pictures of labels and price tags photographed from clothes shop windows. These two sub-corpora make up the CONDIVport corpus, which is still under construction (Soares da Silva 2008b). This corpus is structured according to geographical, diachronic and stylistic variables and has, at present, an extension of 4 million tokens from the formal register (used in sports newspapers and fashion magazines) and 15 million tokens from the informal register (of Internet football chats and clothes labels). The CONDIVport corpus is partly available on the Linguateca website www.linguateca.pt/ACDC (a distributed resource center for language technology for Portuguese; Santos and Sarmento 2003, Santos 2009).

The analysis was carried out for 21 sets of synonymous terms (or onomasiological *profiles*) from the lexical field of football, which means that a total number of 183 terms were studied in a database containing 90,202 observations of these terms used in sports newspapers and 143,946 observations of their use in Internet chats. The analysis is also comprised of 22 onomasiological profiles of clothing items for men (M) and women (F), which means that 264 terms were studied in a database compiling 12,451 observations of their use in fashion magazines and 3,240 observations of their use in labels and price tags pictured from clothes shops. All the profiles including their denotational synonyms are listed in the appendix (terms with a strong popular mark were excluded to avoid inflating differences).[2] The name of each profile is translated into English. The profiles for football are: BACK, BALL, COACH, CORNER, DRIBBLING, FORWARD, FOUL, FREE KICK, GOAL$_1$, GOAL$_2$, GOALKEEPER, MATCH, MIDFIELDER, OFFSIDE, PENALTY, REFEREE, ASSISTANT REFEREE, SHOT/KICK, SHOT/PLAYING, TEAM, WINGER. The profiles for clothing are: BLOUSE F, CARDIGAN M/F, COAT F, COAT M, DRESS F, JACKET M/F, JACKET (BLOUSON) M/F, JEANS M/F, JUMPER M/F, LEGGINGS F, OVERCOAT M/F, RAINCOAT M/F, SHIRT M, SHORT JACKET F, SHORT JACKET M, SHORT TROUSERS M/F, SKIRT F,

SUIT M, SUIT/OUTFIT F, TAILORED JACKET M/F, TROUSERS M/F, T-SHIRT
M/F.

The quantitative methods used to measure convergence and divergence
as well as other types of distances between EP and BP are uniformity (U)
measures and featural (A) measures. Both were developed by Geeraerts,
Grondelaers, and Speelman (1999). The U measure or "profile-based lin-
guistic uniformity" (Geeraerts 2001; Speelman, Grondelaers, and Geeraerts
2003) is a basic measurement which involves the notions of (formal ono-
masiological) *profile* and *uniformity*. The *onomasiological profile* of a con-
cept in a particular language variety is the set of alternative synonymous
terms used to designate that concept in that language variety, together with
their frequencies. *Uniformity* is a measure for the similarity between the
profiles in the different language varieties. For instance, uniformity be-
tween two samples of data is obtained as follows: if term A occurs 6 times
and term B occurs 4 times in one of the samples, and term A occurs 3 times
and term B occurs 7 times in the other sample, the number of common pairs
that name this concept (7 pairs) constitutes a uniformity of 70%. This result
is obtained by making the sum of the lowest relative frequencies of each
alternative term: 30% of term A (3 uses, coming from the second sample)
and 40% of term B (4 uses, coming from the second sample).

Technically, the uniformity for a concept can be calculated with the fol-
lowing formula (see Geeraerts, Grondelaers and Speelman 1999: 36-64 for
further details on the formulae presented in this section):

$$U_z(Y_1, Y_2) = \sum_{i=1}^{n} \min(F_{z,Y_1}(x_i), F_{z,Y_2}(x_i))$$

The uniformity U for a concept Z between two samples Y_1 and Y_2 equals
the sum of the minima of relative frequencies F of the lexical item x in the
onomasiological profiles for Z in Y_1 and Y_2.

If we wish to investigate more than one concept at a time, uniformity U
is defined as the average of uniformity ratings obtained for isolated con-
cepts, whereas uniformity U' is defined as the *weighted average* of the
relative frequency of each concept in the researched samples. The weighted
uniformity U' is calculated as follows:

$$U'(Y_1, Y_2) = \sum_{i=1}^{n} U_{z_i}(Y_1, Y_2) \cdot G_{z_i}$$

The uniformity U' for a set of concepts Z between two samples Y_1 and Y_2 equals the sum of U-values for Zs weighted by the relative frequencies G of Z within the total set of Zs.

The weighted uniformity U' implies that high frequency concepts have a more outspoken impact on the overall uniformity, whereas the unweighted uniformity U presupposes that all the concepts hold the same status. The relation between Y_1 and Y_2 – in the present study, between EP and BP – is accounted for from a pragmatic and communicative perspective rather than a structural one. Hence, the U' measure is more significant than the U measure. In fact, the attested occurrences of an onomasiological profile are an important factor for calculating the convergence or divergence between language varieties.

Diachronically, convergence and divergence can be quantified through increasing or decreasing uniformity. Synchronically, the greater the distance there is between the standard and substandard registers, the smaller uniformity there is between these two registers. We consider that only statistical differences above 5% are significant, whereas smaller differences are not. This is a rule arbitrarily chosen to account for a statistical margin of error.

Table 1. Uniformity (U) for the FORWARD profile between EP and BP (1950-1970)

FORWARD	P50	B50	U	P70	B70	U
atacante	8,8	36,6		13,6	73,8	
avançado	71,6	0,9		47,4	0,0	
avante	0,0	48,9		0,0	11,0	
dianteiro	19,2	6,8		20,1	0,7	
forward	0,1	5,2		0,0	0,0	
ponta-de-lança	0,3	1,5		19,0	14,5	
			16,9			28,8

Table 1 shows the percentages of the U measure in relation to the onomasiological profile FORWARD in the Portuguese (P) and Brazilian (B) databases between 1950 and 1970 (P50, B50, P70, B70). These percentages equal the sum of the smallest relative frequency for each alternative term, i.e. 8.8 + 0.9 + 0 + 6.8 + 0.1 + 0.3 = 16.9% and 13.6 + 0 + 0 + 0.7 + 0 + 14.5 = 28.8%. The increase in uniformity between EP and BP from 16.9% in the

1950s to 28.8% in the 1970s suggests convergence between both varieties in relation to the FORWARD profile.

Another profile-based uniformity measure consists of calculating uniformity within a single language variety. The *internal* uniformity reaches its highest value when all the speakers, in every circumstance, choose the same lexical item to denote a given concept. The internal uniformity value will decrease the more terms there are competing to denote the same concept, and the more dominant some of these terms become. The internal uniformity measure, or I/I' measure, can be calculated with the following formulae, which are related to the external uniformity formulae presented before.

$$I_Z(Y) = \sum_{i=1}^{n} F_{Z,Y} (x_i)^2$$

The internal uniformity I for a concept Z in the sample Y equals the square sum of relative frequencies F of the lexical item x in the onomasiological profile for Z in Y.

$$I'(Y) = \sum_{i=1}^{n} I_{Zi} (Y).G_{Zi} (Y)$$

The internal uniformity I' for a set of concepts Z in the sample Y equals the sum of I-values for Zs weighted by the relative frequencies G of Z within the total set of Zs in Y.

Table 2. Internal uniformity (I) for the FORWARD profile in EP and BP in the 1950s

FORWARD	P50	I	B50	I
atacante	8,8	77,8	36,6	1340,7
avançado	71,6	5128,8	0,9	0,9
avante	0,0	0,0	48,9	2393,5
dianteiro	19,2	369,2	6,8	45,8
forward	0,1	0,0	5,2	27,4
ponta-de-lança	0,3	0,1	1,5	2,4
		55,8		38,1

The I/I' measure is here used mainly in a comparative perspective. Consider again the example of the FORWARD profile. As Table 2 shows, the inter-

nal uniformity is greater in the 1950s Portuguese database (I 55.8%) than in the 1950s Brazilian database (I 38.1%). This can be explained by the two factors which contribute to determine internal uniformity. First, P50 has a single term which is clearly dominant whereas B50 has two dominant terms. Second, there are more highly frequent alternative terms in B50 than in P50.

The proportion of terms possessing a special feature, or A measure, is given in the following formulae. As with the U and U' measures, the A' measure takes into account the relative frequency of each concept whereas the A measure does not.

$$A_{K,Z}(Y) = \sum_{i=1}^{n} F_{Z,Y}(X_i). W_{Xi}(K)$$

The proportion A of all items x with feature K in the onomasiological profile of a concept Z in the subcorpus Y equals the sum of x's relative frequencies weighted by the membership value W.

$$A'_{K}(Y) = \sum_{i=1}^{n} A_{K,Zi}(Y). G_{Zi}(Y)$$

The proportion A' of all items x with feature K in the subcorpus Y equals the sum of all A-measures, weighted by G, that is the relative frequency of concept Z in Y.

Table 3. The impact of the English loans (A) on the GOAL$_1$ profile in EP and BP in the 1950s

| GOAL$_1$ | P50 | | | B50 | | | W |
	abs	rel	rel*W	abs	rel	rel*W	
bola	109	3,7	0,0	0	0,0	0,0	0
goal	24	0,8	0,8	528	38,8	38,8	1
gol	0	0,0	0,0	111	8,1	4,1	0,5
gôl	0	0,0	0,0	66	4,8	1,9	0,4
golo	1841	61,9	31,0	0	0,0	0,0	0,5
ponto	204	6,9	0,0	26	1,9	0,0	0
tento	795	26,7	0,0	631	46,3	0,0	0
			31,8			44,8	

Assigning a special feature K is not a binary issue, but rather the result of a continuum. If we consider the loanword feature, for example, the highest

score (1) is given to loanwords keeping their original form, and the lowest score (0.25) to strongly adapted terms and loanwords' translations. In Table 3, the English influence on the onomasiological profile GOAL$_1$ in the 1950s is measured, based on the absolute (abs) and relative (rel) frequencies of the alternative terms, the W value of the English loan, and the sum of the relative frequencies of the alternative terms weighted by the membership value W (rel*W), i.e. A (P50) 31.8% and A (B50) 44.8%.

3.   External diachronic analysis: uniformity results

In this section, we will analyze the evolution of the two language varieties from an external perspective, that is, the relation between EP and BP taking each variety as a whole. The internal linguistic factors which may have played a role in the global evolution of the two varieties will be discussed in section 4.

Three questions need to be asked:

(1)   Is there convergence or divergence between EP and BP?
(2)   Does the convergent/divergent trend occur on both sides or mainly in one of them?
(3)   Is uniformity increasing or decreasing within each variety? Is internal uniformity greater in EP or BP?

Uniformity calculations U and U' were used to answer question (1) and (2). Convergence and divergence (question 1) are expressed through the increase and decrease of U/U' values, respectively. The convergent or divergent evolution of one of the varieties towards the other (question 2) is expressed by greater changes of U/U' values from a time period to another and by higher or lower values of U/U' in different time periods rather than in a particular time period. Internal uniformity calculations I and I' were used to answer question (3): the increase of I/I' is indicative of growing internal homogeneity. As mentioned before, in principle, weighted measures (U', I') are more significant than unweighted ones (U, I) and a statistical difference below 5% is not treated as significant.

Table 4 presents the results of uniformity U and weighted uniformity U' for each one of the 21 concepts/profiles of football for the Portuguese variety of Portugal (P) and the Portuguese variety of Brazil (B) in the decades of the 1950s, 1970s and 1990s-2000s; the column on the right shows the

total number of observations of each profile and the absolute total of observations.[3]

Figure 1 systematizes the percentages obtained in the calculation of external (U) and internal (I) uniformity for football terms: the first number of each pair of results represents the unweighted uniformity (U, I) and the second number the weighted uniformity (U', I'); the results displayed on the horizontal lines are the percentages of uniformity (U/U') between EP and BP in the 1950s, 1970s and 2000s, which answer question (1); the results shown on the vertical and diagonal lines are the percentages of U/U' obtained from one time period to another and between the different time periods, also providing the answer to question (2); the results associated to each variety and time period are the percentages of internal uniformity (I/I') and the answer to question (3).

Table 4. U and U' percentages for the 21 football profiles

	P50/B50		P70/B70		P00/B00		Total
	U	U'	U	U'	U	U'	nº
ASSISTANT REFEREE	20,1	0,09	43,4	0,33	18,2	0,18	623
BACK	19,9	0,48	15,2	0,5	34,3	1,33	2791
BALL	61,2	4,99	81,9	7,04	95,0	4,75	6542
COACH	49,8	0,69	73,5	3,43	65,4	7,05	4720
CORNER	0,5	0,01	0,0	0	0,0	0	818
DRIBBLING	73,1	0,22	67,1	0,23	54,5	0,11	251
FORWARD	16,9	0,65	28,8	0,82	10,1	0,39	3238
FOUL	60,2	0,61	93,0	0,64	92,7	0,85	814
FREE KICK	0,0	0	1,0	0,01	6,1	0,06	675
GOAL$_1$	42,5	4,79	93,8	12,2	94,1	12,9	11294
GOAL$_2$	28,9	1,35	12,2	0,42	7,5	0,17	3250
GOALKEEPER	8,0	0,21	1,0	0,03	0,0	0	2332
MATCH	54,6	14,1	75,5	17,4	73,0	15,9	21502
MIDFIELDER	48,3	1,22	19,7	0,13	0,9	0,03	2004
OFFSIDE	2,8	0,01	0,0	0	0,0	0	395
PENALTY	27,3	0,41	1,9	0,02	0,6	0,01	1450
REFEREE	43,9	1,76	22,0	0,82	90,2	2,85	3310
SHOT/KICK	47,8	0,9	18,6	0,23	6,4	0,05	1211
SHOT/PLAYING	91,3	3,97	69,4	3,84	52,7	2,2	4140
TEAM	37,3	7,19	33,5	7,04	41,8	7,83	17642
WINGER	6,5	0,14	2,6	0,04	29,3	0,08	1200
Total	35,3	**43,8**	35,9	**55,2**	36,8	**56,8**	90202

As regards question (1), the percentages obtained for U' are suggestive of convergence between the 1950s and 1970s and subsequent stability, but there are no significant changes in U percentages. U' rises 11.39% between

the 1950s and the 1970s (from 43.78% to 55.17%) and increases by 1.59%
only between the 1970s and 2000s, while the unweighted U remains fairly
even. This means that convergence is found at the level of the most fre-
quent concepts (see Table 4). At the same time, the results show a great
distance between the two varieties along the three time periods. In fact, the
results show uniformity values between 44% and 57%, which means that
half of the sample corresponds to differences between the varieties.

These results, therefore, do not confirm the divergence expectation between
EP and BP. There seems to be a convergence pattern in the first two pe-
riods, but only at the level of the weighted measure. The convergence pat-
tern is not very clear, since the difference between the percentages from
both periods is not high.

It should be mentioned that all the phonetic and graphic variants were
considered as alternative terms of the onomasiological profile in question.
For instance, the original form of the loanword *goal* and its adaptation
gol(o) or the terms *penalty*, *penalti* and *pênalti* were considered alternative
terms in relation to the profiles GOAL and PENALTY, respectively. However,
golo (P) and *gol* (B), as well as *chuto* (P) and *chute* (B), were not split into
alternative terms, because they are intrinsic to their respective national va-
riety.
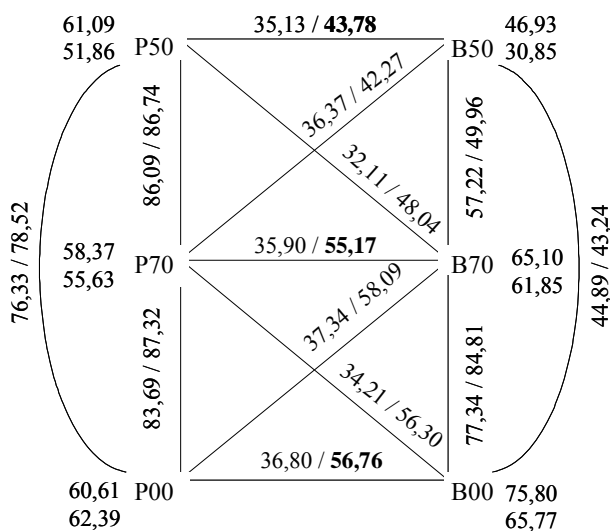
Figure 1. Uniformity results for football terms

Another interesting question is to know what happens when other types of calculations are used instead. The results of four alternative calculations are the following:

— when we separate *golo/gol* and *chuto/chute,* the result points towards a stable situation:
U'(P50,B50) 42.16% $\cong$ U'(P70,B70) 43.42% $\cong$ U'(P00,B00) 43.83%
— when GOAL$_1$ is excluded from the calculation, the result remains identical to the first one, i.e. convergence between 1950 and 1970:
U'(P50,B50) 43.81% < U'(P70,B70) 49.33% $\cong$ U'(P00,B00) 50.97%
— when all the phonetic and graphic variants are taken into consideration together as one term in the calculation, i.e. *penalty/penalty/pênalti* as one term, and *goal/gol(o)* as another term, the result still suggests convergence between the 1950s and the 1970s for U':
U'(P50,B50) 48.09% < U'(P70,B70) 56.22% $\cong$ U'(P00,B00) 57.61%
— excluding the most asymmetric profiles, i.e. the more frequent profiles (MATCH, TEAM and GOAL$_1$) and the less frequent ones (OFFSIDE and DRIBBLING), the result shows a stable situation:
U'(P50,B50) 40.82% $\cong$ U'(P70,B70) 43.53% $\cong$ U'(P00,B00) 44.24%

Some problems arise from the above results. The first alternative calculation suggests that the profile GOAL$_1$ — which is the third most frequent profile and represents 13% on average of the overall quotations (see Table 4) – is responsible for modifying the results. A similar stable situation is evidenced by the result of the last alternative calculation. This means that the inclusion or exclusion of a variant or of a single concept may change the picture entirely, which may pose some problems, particularly since the studied concepts were picked by hand and do not represent the entire lexical field. However, the remainder of the calculations still indicates convergence between 1950 and 1970. (Interestingly, the result is the same whether or not the phonetic and graphic variants are separated). Furthermore, the concepts studied are representative of the lexical field of football and there is a balance between the more frequent and the less frequent concepts.

The question of the preference for the weighted measure may be more problematic. Given the alternative calculations, we might in fact question whether the unweighted measure is not being given less attention than it may actually deserve. We reiterate that, for the present study, the pragmatic perspective (which integrates the differences in frequency of the concepts studied) is more important than the structural perspective (which attributes the same weight equally to every concept). Furthermore, it was observed

that the low frequency of OFFSIDE and DRIBBLING, for instance, is due to the low occurrence of these concepts in texts related to football. For this reason, and because the concepts studied are common, the unweighted calculation will continue to be used.

Another issue in our discussion is the behavior of the profiles, i.e. which profiles behave differently from the overall convergence tendency, and whether they have anything in common with each other. The profiles with divergent results are GOAL$_2$, DRIBBLING, SHOT/PLAYING, MIDFIELDER, SHOT/KICK and, to a lesser extent, FORWARD, OFFSIDE, PENALTY and GOALKEEPER, which do not seem to exhibit common characteristics such as to constitute a subgroup of their own. The only common characteristic is the fact that some of the profiles relate to less frequent concepts, as Table 4 shows (we will come back to this question at the end of this section).

In conclusion, the results seem to indicate a slight convergence between EP and BP between 1950 and 1970. However, this is a restricted convergence. The evidence comes from the fact that the results change according to the inclusion/exclusion of certain concepts, and to the evolutionary stability given by the unweighted measure.

As regards question (2), the percentages obtained for U and U' shown on the vertical lines in Figure 1 indicate that many more changes occurred in the case of BP over a long term period (between 1950 and 2000) and also in the first time period (from the 1950s to 1970). Furthermore, BP seems to have got closer to EP between 1950 and 1970 (see the diagonal line): B70 not only came closer to P70 but also to P50, namely U' (B70, P50) 48.04% is greater than U' (B50, P50) 43.78%.

What the results clearly show is that internal evolution is stronger and faster in BP than in EP. We could infer from this result that the approach happens mainly in BP. This interpretation is problematic, though. On the one hand, the fact that BP undergoes many changes may suggest that it is recovering from a standardization underdevelopment in vocabulary related to football (or in vocabulary in general), or that it wishes to conform to the standard EP variety. On the other hand, the influence of Brazilian football is well known and many Brazilian players and technicians have come to Portugal. The expectation of the growing influence of BP over EP, particularly in relation to football, as mentioned earlier, would mean, on the contrary, a greater shifting of EP towards BP, which is not confirmed by the results presented in Figure 1. In fact, the difference between U' (P00, B70) 58.09% and U' (P70, B70) 55.17% is not high enough to warrant the consideration that EP is moving towards BP.

What we can safely state now is that BP has clearly undergone more changes than EP, which does not necessarily mean that BP is nearing any closer to EP, in the sense of BP conforming to EP. In order to interpret the results more thoroughly, the features of the terms studied need to be further analyzed. As we will see in the next section, there is a factor which may partly explain why BP exhibits greater changes, as well as the apparent approaching of BP towards EP in the first two periods. This factor is the introduction of loanwords which have had a greater influence on BP.

As for question (3), the percentages of I and I' are indicative of a great increase in internal uniformity in BP, mainly between 1950 and 1970, whereas minor change is found in EP. I and I' measures are a lot lower for B50 than P50; but happen to be higher for B70 and B00 as compared to P70 and P00.

BP undergoes global and intermediate changes more clearly than EP. That change is directed towards a strong increase in internal uniformity. This may not always be the case, since if a variety changes a lot, such change may either increase or decrease internal uniformity. The increase in internal uniformity, which is stronger in BP, is due to the increase in internal homogeneity derived from two factors: i) the decrease in competition between onomasiological alternatives in the majority of the profiles in favor of the dominance of a single term (sometimes two terms), and ii) a reduction in the number of alternative terms. It is BP which exhibits a greater decrease in the number of alternative terms; only in 2 (BACK and MID-FIELDER) of the 21 onomasiological profiles are there more synonyms in BP00 than in EP00; in the rest of the profiles there are either fewer synonyms in BP00 (11 profiles) or the same number (8 profiles); there are 4 profiles with only 1 term in BP00 (FOUL, OFFSIDE, GOALKEEPER and FREE KICK), which is not observed in EP00; and finally, within the set of the 21 profiles and respective 183 terms, 93 are from EP00 against 71 in BP00.

A growing internal homogeneity in the field of football is thus observed in both varieties between 1950 and 2000, but the tendency is higher in BP than in EP. At present, this homogeneity has similar percentages in both varieties, although BP exhibits a higher internal uniformity at the level of the unweighted measure – I (B00) 75.80% > I (P00) 60.61%. We can interpret this increase in internal uniformity as an indicator of standardization, in the sense that standardization can be seen as a process which reduces the internal formal onomasiological variation. This interpretation is set in the context of the growing popularity and globalization of football over the last 60 years. The more popular and globalized football becomes, the more standardized its vocabulary tends to be. However, the association between

internal uniformization and standardization is not crucial, because a standardized linguistic situation does not necessarily imply a stable internal uniformity. There can be onomasiological changes within a standardized linguistic situation; if this were not the case, linguistic change would be impossible in a standardized linguistic context. The results in Figure 1 show that there are also fluctuations in internal uniformity in EP, especially between 1970 and 2000. The theoretical hypothesis that EP reached a relative standardization situation faster does not mean that within EP changes of onomasiological preferences ceased to take place. Nonetheless, fewer changes are clearly observed in EP than in BP.

Another reason not to use internal uniformity necessarily as an indicator of standardization is due to the fact that it is not possible to know how much internal variation is normal or acceptable to consider whether a given linguistic situation is standardized. It is, however, possible to overcome this problem if we take one of the varieties as a reference point, in this case the one which is hoped to exhibit fewer changes – EP. The results allow us to reach, albeit to a certain extent, the conclusion that there are no significant changes of internal uniformity in EP, except between 1970 and 2000 at the level of the weighted measure. However, the same results disallow any other way to take EP as the reference point, namely the idea that the level of internal uniformity of BP is, by default, lower than in EP, precisely because this only happens in the 1950s.

We can thus conclude two things: first there are some aspects in BP in the 1950s related to the terms studied which need further analysis – we shall see this in the next section. Second, the greater increase in internal uniformity in BP than that in EP may be interpreted, within the social-cultural context of football, as increased standardization of Brazilian vocabulary related to football, in comparison with EP. This does not mean, however, that EP has served as a role model for BP.

There is also another question related to the values of internal uniformity. Do the results of I/I' calculations suggest a more recent standardization of the Brazilian variety? There is no simple answer. Three reasons may explain this strong rise in internal uniformity. First, the low values of I/I' for B50 are highly correlated with the huge number of loanwords kept in their original form (see the next section). Second, the popularity and international prestige of Brazilian football rose in the 1960's after two World Cup victories in a row (1958 and 1962). Third, as has already been mentioned, formal onomasiological variation has lost ground in both varieties, and this is more perceptible in BP. All these reasons may possibly indicate a greater standardization of the vocabulary of football in BP than in EP. But

they may, at the same time, or alternatively, suggest a less careful attitude towards stylistic refinement in today's Brazilian press, compared to the Portuguese press.

We will now turn to the results regarding clothing terms. Table 5 presents U/U' values for the 22 onomasiological profiles from the lexical field of clothing. Figure 2 systematizes U/U' and I/I' results related to the three issues analyzed in this section. These results are presented in the same way as the football terms (Figure 1).

Table 5. U and U' percentages for the 22 clothing profiles

	P50/B50		P70/B70		P00/B00		Total
	U	U'	U	U'	U	U'	n°
BLOUSE F	86,9	8,8	79,4	8,0	85,3	7,8	1213
CARDIGAN M/F	87,3	0,4	7,1	0,1	15,5	0,1	97
COAT F	78,6	4,0	41,0	3,0	37,5	3,2	896
COAT M	3,9	0,1	5,1	0,1	22,2	0,1	138
DRESS F	96,2	23,0	80,6	13,2	92,4	16,1	2335
JACKET M/F	36,9	0,5	88,7	1,9	35,3	0,4	198
JACKET (BLOUSON) M/F	15,4	0,1	98,4	1,8	76,9	1,6	193
JEANS M/F	0,0	0,0	0,0	0,0	77,5	2,9	180
JUMPER M/F	36,8	0,7	34,4	2,1	7,1	0,4	573
LEGGINGS F	0,0	0,0	0,0	0,0	81,8	0,2	12
OUTFIT F	61,5	6,6	79,3	5,9	23,9	0,9	849
OVERCOAT M/F	39,7	2,7	56,7	3,3	71,8	2,5	642
RAINCOAT M/F	60,3	1,2	65,1	1,5	34,3	0,3	207
SHIRT M	100,0	1,3	83,3	0,7	97,6	0,9	128
SHORT JACKET F	69,3	1,3	66,0	1,2	79,2	0,9	195
SHORT JACKET M	85,0	0,6	0,0	0,0	100,0	0,1	39
SHORT TROUSERS M/F	68,2	1,3	67,9	2,6	19,9	0,9	454
SKIRT F	99,8	21,9	90,5	16,2	86,1	15,0	2337
SUIT M	0,0	0,0	0,0	0,0	0,0	0,0	88
TAILORED JACKET M/F	79,1	1,9	80,0	0,1	56,3	0,2	106
TROUSERS M/F	66,4	2,0	25,5	2,9	13,6	2,0	1327
T-SHIRT M/F	0,0	0,0	65,3	1,0	17,7	0,6	244
Total	61,6	**78,4**	58,7	**65,7**	51,4	**57,1**	12451

As for question (1), U and U' percentages give evidence of a divergent trend that has emerged through time: U' decreases 12.88% between 1950 and 1970 (from 78.80% to 65.92%) and 8.81% between 1970 and 2000 (from 65.92% to 57.11%); generally, U' decreases 21.69% throughout the whole period.

As an alternative calculation, all the phonetic and graphic variants such as *blouse* and *blusa* (blouse) or *blaizer* and *blêizer* (COAT), for example, are not separated. The result still points towards a divergent evolution:

U'(P50,B50) 82.63% > U'(P70,B70) 77.10% > U'(P00,B00) 71.91%

Compared to football terms, clothing terms not only exhibit an opposite global result but they also show a much clearer evolutionary relation between the two varieties. In fact, both the percentages related to weighted uniformity U' and unweighted uniformity U indicate that there is divergence, which is more expressive throughout the period under study.

As regards question (2), the percentages of U and U' shown on the vertical and diagonal lines are indicative of symmetry in the evolution trends of both varieties (global changes and intermediate changes are identical) and show that divergence occurs on both sides, in both time periods.
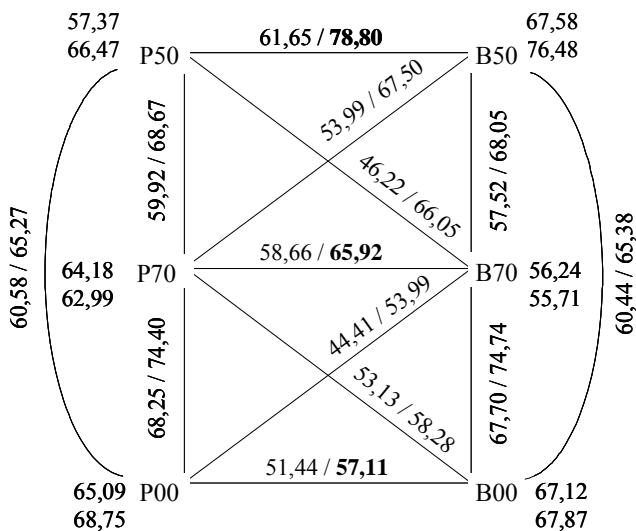
Figure 2. Uniformity results for clothing terms

The percentages on the vertical lines show that the changes are not stronger or faster in one variety than in the other, but, rather, that they are identical in the two varieties not only between 1950 and 2000 but also between the interim periods. As a matter of fact, the differences in the percentages when comparing EP to BP are never significant (they are always below the 5%

margin of error). This suggests that there is a real internal distancing process which takes place in both varieties. The same is observed with the percentages on the diagonal lines: the uniformity between the varieties in two different periods (P70 and B50, B70 and P50, P00 and B70, B00 and P70) is always inferior to the uniformity between the varieties in the preceding period (P50 and B50, etc.). This means that not only does the Brazilian variety move away from the European variety, but the latter also moves away from the former. The hypothesis of growing symmetrical divergence would obviously require that more intermediate periods be available for testing (between 1950 and 1970, and 1970 and 2000). Still, an intermediate point (1970) is available, which indicates a growing distance between the two varieties. Considering that the difference of weighted uniformity between 1970 and 2000 is lower than the difference between 1950 and 1970 (8.81% versus 12.88%), we can thus also hypothesize that the divergent movement between the two varieties may have slowed down slightly in the last decades.

As for question (3), I and I' percentages show two things. First, there is a consistent evolutionary pattern: the internal uniformity decreases between 1950 and 1970 (except in EP at the level of the unweighted measure) and increases between 1970 and 2000 in both varieties. The fluctuations in internal uniformity may be related to a process of lexical renovation with regard to fashion, particularly clothing. This will be further analysed in the next section. Contrary to football terms, it is more difficult to find standardization tendencies in relation to clothing terms, even if the period when there is an increase in internal uniformity (1970 and 2000) is taken into consideration. Moreover, the internal homogeneity between the first and last period (1950 and 2000) undergoes some changes, decreasing in BP at the level of the weighted measure. Second, I and I' percentages are evidence of greater changes in BP than in EP, especially between the intermediate periods (long term changes are also observed but only at the level of the weighted uniformity). There is also no clear evidence in this case that one of the varieties may constitute a reference point for the other. In fact, the comparison of internal uniformity between the two varieties reveals different results: in the first period, the internal uniformity is greater in BP than in EP, in the second period, it is lower in BP than in EP, and in the last period it is identical in both varieties.

The results of the external diachronic analysis can be summarized as follows:

— divergence is found in the vocabulary of clothing and restricted convergence is found in the vocabulary of football: the hypothesis of divergence is confirmed in the field of clothing but not in the field of football;
— internal changes occur on a bigger scale in the Brazilian variety in both lexical fields: BP changes more than EP;
— the most significant changes took place between 1950 and 1970;
— an orientation of one variety in relation to the other does not seem to exist (neither approaching nor moving apart);
— the evolutionary situation is more direct and clearer as far as clothing terms are concerned than football terms.

We have thus far only had a few indirect hints about the direction taken by these changes. It is therefore important to study possible correlations between the results obtained until now and the special features of the lexical items under analysis.

First of all, we will examine the distribution of uniformity values U/U' for the different onomasiological profiles of each lexical field in the three periods studied. The purpose is to find out if the slight global convergent tendency observed for football terms and the clear global divergent tendency observed for clothing terms are in any way reflected in the distribution of the uniformity percentages for the respective profiles. In short we want to answer the following questions: do the uniformity values of the individual profiles follow the respective global tendency or are there different uniformity values beyond the global tendency? We will calculate the standard deviation for each of the three moments. Standard deviation is defined as the measure of the degree of variation of the whole set of results: the higher the standard variation is, the greater the differences between the results will be. Standard variation in uniformity values for the set of profiles of each lexical field are as follows:

football: 25.24 (1950); 33.87 (1970); 35.96 (2000);
clothing: 36.36 (1950); 35.19 (1970); 33.67 (2000).

These high values of standard deviation show that the homogeneity between the profiles of each lexical field is not great. In each of the periods studied significant fluctuations between the individual uniformity values are observed, that is, there are always profiles with a uniformity degree much higher than the average and profiles with a uniformity degree much lower than the average. Furthermore, the global convergence tendency observed in football terms is not reflected by the decrease in standard devia-

tion. On the contrary, there is an increase in standard deviation. Similarly, the global divergence tendency observed in clothing terms is not reflected by an increase in standard deviation either. There is, indeed, a decrease in standard deviation.

The analysis of the distribution of the individual uniformity values also shows another aspect: the profiles that best reflect the global evolution trend are those with highest weighted uniformity U' values, i.e., the most frequent ones (in the case of clothing, less frequent profiles also follow the global divergent trend). The results are as follows (note that U' 5% corresponds to U 50%):

for football:
— U' is equal or greater than 5% (7 profiles):
   U'(P50, B50) 50.47% < U'(P70,B70) 65.02% ≅ U'(P00,B00) 69.15%
— U' is lower than 5% (14 profiles):
   U'(P50,B50) 24.37% > U'(P70,B70) 16.70% ≅ U'(P00,B00) 14.40%

for clothing:
— U' is equal or greater than 5% (7 profiles):
   U'(P50,B50) 84.56% > U'(P70,B70) 68.79% > U'(P00,B00) 63.68%
— U' is lower than 5% (15 profiles):
   U'(P50,B50) 51.29% < U'(P70,B70) 57.73% > U'(P00,B00) 37.89%

The comparison between the two lexical fields shows more clearly that the evolutionary tendency between EP and BP is more homogeneous with regard to clothing terms than to football terms. In the former, the most frequent concepts as well as the least frequent terms reflect the global divergent tendency.

4. Internal diachronic analysis: item-related results

We will now examine the impact of two groups of features of the selected items on the global evolution trend:

(1) endo-/exogenousness: we aim to know whether a convergent/divergent evolution is determined by the exogenous/endogenous orientation of one of the varieties towards the other during the converging/diverging process, i.e. if this occurs through the adoption of exogenous terms and giving up of endogenous terms, or the opposite

(2) foreign influence (loanwords): to assess the impact of English and French influence on the global convergent/divergent evolution.

Endo-/exogenousness is mathematically derived from the corpus: *endogenous* words are typical terms found in a sub-corpus (they occur more frequently in this sub-corpus) and *exogenous* words are typical terms from the rest of the corpus. Hence, a word which is endogenous in a language variety will be exogenous in the other language variety and vice-versa. There is yet a third feature – normative influence – which could be considered, so as to evaluate the consequences of linguistic propaganda on the global convergent/divergent evolution. At this stage, unfortunately, it cannot be assessed due to lack of data.

4.1. Endo-/exogenousness

This parameter involves the calculation of the proportion of items with 'binational' (AUNI), 'endogenous' (AENDO) and 'exogenous' (AEXO) features. Featural measures A and A' mentioned in section 2 are used to calculate the proportion of items exhibiting these features. Before A and A' are calculated, ENDO, EXO and UNI values in relation to the studied terms must be measured. The endo-/exogenousness values of a term X – denoting a concept Z – are calculated on the basis of the differential influence of X over the relation between the onomasiological profiles of the concept Z in Y_1 e Y_2 samples under comparison. Consider the following example: if X's relative frequency is 30% in Y_1 and 0% in Y_2, then X's differential value is maximal, that is, 1 (in a 0–1 scale) or 100%. Also in this example, X's ENDO value is 1 in Y_1 and X's EXO value is 0 in Y_1. By definition, X's endogenousness value in Y_1 and X's exogenousness value in Y_2 are the same, and the sum of the X's endogenousness and exogenousness values equals its differential value. On the other hand, X's UNI value corresponds to the difference between 1 (the maximal value in the 0–1 scale) and its differential value. While ENDO e EXO measure the stronger or weaker position of X in Y_1 in comparison with Y_2, respectively, UNI measures X's contribution to the uniformity between Y_1 and Y_2 (see Geeraerts, Grondelaers and Speelman 1999: 49-53 for technical details). A direct relation between the proportion of the terms with a certain UNI value and uniformity measure U used in the previous section can, therefore, be posited.

   We can use the endo-/exogenousness measures and relate them to the global evolutionary patterns described above through uniformity measures.

For instance, a convergent evolution implies an increase in AUNI. It is necessary, nonetheless, to check if the increase in AUNI is due to AENDO, AEXO or both. In other words, a convergent evolution may originate either in the exclusion of endogenous terms or in the inclusion of exogenous terms. Mathematically, we could then state that the convergent evolution implies an increase in binational and exogenous terms and a decrease in endogenous terms; conversely, a divergent evolution involves a decrease in binational and exogenous terms and an increase in endogenous terms.

Table 6 presents the percentages of endo-/exogenousness in the corpus of football in the Portuguese (P) and Brazilian (B) variety in the three periods studied (only the values of the weighted A' measure are indicated). Three results stand out: a marked increase of A'UNI between 1950 and 1970 on the Brazilian side (from 38.69% to 55.82%); a decrease of A'ENDO between 1950 and 1970 which appears to be stronger in Brazil from the 50's to the 70's (from 51.78% to 38.62%); and greater changes of A'EXO in Brazil. These changes are consistent with the convergent trend previously observed (between 1950 and 1970) and confirm that the Brazilian variety is subject to greater change than the European variety.

Table 6. Evolution of bi-national, endogenous and exogenous football terms

A'UNIp50,b50(P50) < A'UNIp70,b70(P70) ≅ A'UNIp00,b00(P00)		
48,47	54,98	58,01
A'EXOp50,b50(P50) ≅ A'EXOp70,b70(P70) ≅ A'EXOp00,b00(P00)		
4,09	4,99	4,24
A'ENDOp50,b50(P50) > A'ENDOp70,b70(P70) ≅ A'ENDOp00,b00(P00)		
47,44	40,03	37,74
A'UNIb50,p50(B50) < A'UNIb70,p70(B70) ≅ A'UNIb00,p00(B00)		
38,69	55,82	55,84
A'EXOb50,p50(B50) ≅ A'EXOb70,p70(B70) ≅ A'EXOb00,p00(B00)		
9,53	5,56	5,93
A'ENDOb50,p50(B50) > A'ENDOb70,p70(B70) ≅ A'ENDOb00,p00(B00)		
51,78	38,62	38,23

Nonetheless, the calculation of endo-/exogenousness may not reflect reality. The reason for this is that we simply cannot take into account AUNI, AENDO e AEXO from period to period, because what is endogenous or exogenous can produce different results in each period. The most important measure is the one that allows us to understand what happened during the

1970s to the terms that were endogenous, exogenous or binational in the 1950s and what happened to them in 2000 compared with 1970. The results are given in Table 7.

Table 7. Sequential evolution of bi-national, endogenous and exogenous football terms

A'UNIp50,b50(P50) ≅ A'UNIp50,b50(P70)			A'UNIb50,p50(B50) < A'UNIb50,p50(B70)	
48,47	49,98		38,69	45,32
A'EXOp50,b50(P50) ≅ A'EXOp50,b50(P70)			A'EXOb50,p50(B50) < A'EXOb50,p50(B70)	
4,09	4,86		9,53	18,54
A'ENDOp50,b50(P50)		≅	A'ENDOb50,p50(B50) > A'ENDOb50,p50(B70)	
A'ENDOp50,b50(P70)			51,78	34,50
47,44	44,65		A'UNIb70,p70(B70) ≅ A'UNIb70,p70(B00)	
A'UNIp70,b70(P70) < A'UNIp70,b70(P00)			55,82	55,10
54,98	60,10		A'EXOb70,p70(B70) ≅ A'EXOb70,p70(B00)	
A'EXOp70,b70(P70) ≅ A'EXOp70,b70(P00)			5,56	8,00
4,99	5,56		A'ENDOb70,p70(B70) ≅ A'ENDOb70,p70(B00)	
A'ENDOp70,b70(P70)		>	38,62	35,48
A'ENDOp70,b70(P00)				
40,03	33,72			

The upper part of Table 7 shows that from the 1950s to the 1970s the Brazilian variety (on the right) undergoes more changes than its European counterpart (on the left). All three changes regarding the proportion of UNI, EXO and ENDO terms between 1950 and 1970 are significant in BP (i.e. above 5%). None of these changes are significant in EP, though. More specifically, between 1950 and 1970, BP shows an increase in binational and exogenous terms and a decrease in endogenous terms. In EP, in contrast, the proportion of binational, exogenous and endogenous terms remains identical in the same period. The evolution of the binational, exogenous and endogenous terms just described confirms that BP moves closer to EP, as proposed earlier (see the diagonal line between B70 and P50 in Figure 1). The factor causing the varieties to draw closer together is, for now, still unknown. Furthermore, despite confirmation that BP is approaching towards EP, this does not mean that EP was the role model for BP. It is BP which moves, thus becoming closer to EP in the 1970s, but this does not necessarily imply an orientation on the part of BP towards conforming to

EP. Looking at the lower part of Table 7, we can see an increase of A'UNI and a decrease of A'ENDO in EP from 1970 to 2000. This may suggest movement towards BP, although the fact that the increase of A'EXO is higher in BP than in EP poses a problem for such a suggestion.

An important question is to determine what is more significant: the increase in exogenous terms or the decrease in endogenous terms. If we compare the increase of the EXO proportion and the decrease of the ENDO proportion, we notice that the growth of the EXO proportion is, in the majority of cases, more marked than the decrease of the ENDO proportion in BP and EP (except in EP from 1970 to 2000), and this trend is even more noticeable in BP than EP. All this means that the two varieties (BP more than EP) adopted more exogenous terms than they gave up endogenous terms during these time periods. In relation to the apparent shifting of BP towards EP from 1950 to 1970, this would imply an exogenous approach, which appears to suggest a movement oriented towards EP. However, we have to analyze other internal features so as to offer a more considered interpretation of this nearing of the varieties.

Table 8. Evolution of bi-national, endogenous and exogenous clothing terms

A'UNIp50,b50(P50) > A'UNIp70,b70(P70) > A'UNIp00,b00(P00)		
75,76	67,10	57,78
A'EXOp50,b50(P50) $\cong$ A'EXOp70,b70(P70) $\cong$ A'EXOp00,b00(P00)		
3,31	3,61	4,50
A'ENDOp50,b50(P50) < A'ENDOp70,b70(P70) < A'ENDOp00,b00(P00)		
20,93	29,30	37,72
A'UNIb50,p50(B50) > A'UNIb70,p70(B70) > A'UNIb00,p00(B00)		
82,20	65,50	54,84
A'EXOb50,p50(B50) $\cong$ A'EXOb70,p70(B70) $\cong$ A'EXOb00,p00(B00)		
1,83	4,59	4,74
A'ENDOb50,p50(B50) < A'ENDOb70,p70(B70) < A'ENDOb00,p00(B00)		
15,96	29,91	40,42

Let's now look at clothing terms. Table 8 shows that the global divergence observed earlier is associated with two internal changes: one is the decrease in binational terms; the other change, having a greater impact than the previous one, is the increasing number of endogenous terms on both sides – 16.79% in EP (from 20.93% to 37.72%) and 24.46% in BP (from 15.96% to 40.42%). Another conclusion drawn from Table 8 is that the Brazilian

variety changes more than the European variety, especially in the period
between the 1950s to the 1970s.

In Table 9, the results of the appropriate sequential calculation are pre-
sented. Three facts can be inferred from the percentages. First, we notice
that the percentage of common vocabulary (binational terms) increasingly
falls in the 1950s and 1970s, which is consistent with the global divergent
trend. Second, a strange phenomenon occurs at this point, especially be-
tween the 1950s and the 1970s. While EP discards part of its endogenous
terms (from 20.93% to 13.69%) and takes on exogenous terms (from 3.31%
to 9.41%), at the same time BP adopts more exogenous terms (from 1.83%
to 7.64%). One would expect that this would translate into a greater un-
iformity between the two varieties in the 1970s. But surprisingly, uniformi-
ty decreases considerably, as can be seen in Figure 2. We may be able to
explain this if we take into account the third fact: a growing number of
endogenous terms appear in both varieties during the 1970s: 15.61% more
in EP (29.30% minus 13.69%) and 14.7% more in BP (29.91% minus
15.21%). It is precisely the increase in endogenousness which explains the
divergent evolution observed on both sides. In 2000, the percentage of new
endogenous terms in EP increased by 8.42%, of which 4.41% are inherited
from 1970; whereas BP took on 10.51% of new endogenous terms in 2000,
of which 1.03% are from the 1970s. These results still corroborate the di-
vergent evolution in both varieties. This recent endogenousness is attri-
buted to the entry of new fashion terms and new uses given to existing
terms.

Table 9. Sequential evolution of bi-national, endogenous and exogenous clothing
terms

A'UNIp50,b50(P50) > A'UNIp50,b50(P70)	A'UNIb50,p50(B50) > A'UNIb50,p50(B70)
75,76 65,15	82,20 61,23
A'EXOp50,b50(P50) < A'EXOp50,b50(P70)	A'EXOb50,p50(B50) < A'EXOb50,p50(B70)
3,31 9,41	1,83 7,64
A'ENDOp50,b50(P50) > A'ENDOp50,b50(P70)	A'ENDOb50,p50(B50) $\cong$ A'ENDOb50,p50(B70)
20,93 13,69	15,96 15,21
A'UNIp70,b70(P70) > A'UNIp70,b70(P00)	A'UNIb70,p70(B70) > A'UNIb70,p70(B00)
67,10 52,86	65,50 57,62
A'EXOp70,b70(P70) $\cong$ A'EXOp70,b70(P00)	A'EXOb70,p70(B70) $\cong$ A'EXOb70,p70(B00)
3,61 7,41	4,59 7,55
A'ENDOp70,b70(P70) $\cong$ A'ENDOp70,b70(P00)	A'ENDOb70,p70(B70) $\cong$ A'ENDOb70,p70(B00)
29,30 33,71	29,91 30,94

As regards the percentage variation of endogenousness and exogenousness in the field of clothing in the previous time period, we observe that the changing number of exogenous terms is always larger than the changing proportion of endogenous terms. But if we compare the percentages of endogenousness and exogenousness linearly (defined according to each time period), then the variation of endogenousness values is almost always greater in terms of weighted measurement. This is consistent with the global divergent evolution observed in both varieties.

4.2. Loanwords

We will now examine the impact of loanwords on the global evolution, calculating (using the A/A' measures described in section 2) the proportion of terms with feature 'English', 'French' or 'loan' (regardless of the origin) within the onomasiological profile of a selected concept and then for all the concepts included in the analysis of the samples of both varieties. In the vocabulary of football, foreign loanwords are distributed into two categories: English loans and loans in general (including Spanish, Italian and French loans). In the vocabulary of clothing, foreign borrowings are divided into three categories: French loans, English loans and loans in general. In section 2, it was mentioned that when measuring the weight of loanwords in the onomasiological profile of a given concept, the highest score (1) was given to loanwords keeping their original form and the lowest score (0.25) to strongly-adapted terms and loanword translations.

Tables 10 and 11 show the percentage results obtained for English loans (A'$_{Engl}$), French loans (A'$_{Fr}$) and for all the foreign loanwords in general (A'$_{loan}$) in the Portuguese (P) and Brazilian (B) varieties, in the three periods under study. Only values in relation to weighted measure A' are presented.

As regards the corpus of football, the influence of English borrowings and other loans is clearly stronger in BP than in EP in all the periods studied. Crucially, the huge percentage difference between B50 and P50 (the number of English loans in BP is twice as large as in EP: 18.0% vs. 7.1%) results from a larger number and a higher frequency of foreign borrowings that keep their original form in BP. This is the case with *referee, forward, back, team, foul, goal, keeper, match, half, shoot, corner*, for instance, which are in the majority of cases absent from the European Portuguese texts. This internal feature contributes significantly to a considerable dis-

tance between the two varieties in the 1950s. As Figure 1 shows, the un-
iformity between EP and BP in the 1950s is only 43.78% (U' measure). As
a matter of fact, if we do not separate the adapted and non-adapted variants
of a foreign borrowing in the calculation, that is, if we take, as an example,
penalty and *penalty* as one term, and *goal* and *gol*(*o*) as another term, U'
increases by nearly 5% in 1950 (from 43.78% to 48.12%) and, consequent-
ly, comes closer to the values of U' in 1970 (55.17%) and 2000 (56.76%).
Furthermore, the adaptation of English borrowings and their substitution by
vernacular terms results in a decrease in formal onomasiological variation.
This means that the global uniformity between the two varieties will in-
crease as the football loans become adapted to the language or are replaced
by vernacular terms.

Table 10. Loanwords in the corpus of football terms

A'$_{Engl}$ (P50)	7,1%	18,0%	A'$_{Engl}$ (B50)
A'$_{Engl}$ (P70)	9,8%	17,1%	A'$_{Engl}$ (B70)
A'$_{Engl}$ (P00)	10,2%	16,2%	A'$_{Engl}$ (B00)
A'$_{loan}$ (P50)	13,9%	23,5%	A'$_{loan}$ (B50)
A'$_{loan}$ (P70)	17,9%	22,8%	A'$_{loan}$ (B70)
A'$_{loan}$ (P00)	18,5%	23,3%	A'$_{loan}$ (B00)

Table 11. Loanwords in the corpus of clothing terms

A'$_{Fr}$ (P50)	17,6%	18,5%	A'$_{Fr}$ (B50)
A'$_{Fr}$ (P70)	15,9%	18,1%	A'$_{Fr}$ (B70)
A'$_{Fr}$ (P00)	10,2%	7,9%	A'$_{Fr}$ (B00)
A'$_{Engl}$ (P50)	3,3%	4,2%	A'$_{Engl}$ (B50)
A'$_{Engl}$ (P70)	5,8%	7,6%	A'$_{Engl}$ (B70)
A'$_{Engl}$ (P00)	16,9%	17,0%	A'$_{Engl}$ (B00)
A'$_{loan}$ (P50)	22,4%	23,8%	A'$_{loan}$ (B50)
A'$_{loan}$ (P70)	22,1%	26,7%	A'$_{loan}$ (B70)
A'$_{loan}$ (P00)	28,2%	24,9%	A'$_{loan}$ (B00)

This is perhaps the main factor influencing the global convergence between
the two varieties, and more specifically, the approaching of the Brazilian
variety to the European variety from the 1950s to the 1970s. In fact, in the

1950s there are many more loanwords and non-adapted forms in BP than in EP, and, as we can see in Table 12, many more loans are consistently adapted through time and translated in BP than in EP. The consequence of these changes in BP, stronger between the 1950s and the 1970s, is the way BP draws closer to EP, reflected in the evolution of loanwords rather than in the conformation of one variety in relation to the other.

Table 12. Adaptations/translations of English borrowings in the football corpus

$A'_{Engl.adapt}$ (P50)	6,0%	2,8%	$A'_{Engl.adapt}$ (B50)
$A'_{Engl.adapt}$ (P70)	7,9%	16,9%	$A'_{Engl.adapt}$ (B70)
$A'_{Engl.adapt}$ (P00)	8,9%	16,0%	$A'_{Engl.adapt}$ (B00)

Table 13. Evolution of some English loans in the football corpus

	BP	EP
goal-keeper	goleiro (gol + -eiro)	guarda-redes
goal	gol	baliza
corner	córner	pontapé de canto
penalty	pênalti	grande penalidade
back	beque	defesa
shoot	chute	pontapé na bola
offside	impedimento	fora-de-jogo

As for the corpus of clothing, we observe a decrease in French borrowings (stronger in BP) and an increase in English loans in the two varieties, due to well-known sociocultural reasons. The use of foreign loanwords in general is rising in the case of EP and remains quite stable in BP. The influence of foreign loans remains stronger in the Brazilian variety, but the difference between the varieties is weaker in the clothing vocabulary than in the football vocabulary. The relative symmetry of the changes observed in the proportion of loanwords with regard to clothing in the two varieties is overall consistent with the aforementioned global divergent tendency.

As far as both lexical fields are concerned, evidence shows a growing influence of English, particularly in the Brazilian variety (even though a slight weakening of this influence is perceived in the field of football in BP). Evidence also confirms the decreasing influence of French (though the degree of French influence on the vocabulary of football is hardly signifi-

cant) and also reveals that French influence is not weaker in BP than it is in EP. Tables 10 and 11 provide the most significant result, confirming the hypothesis of the Brazilian variety's greater receptivity to loanwords, whether through direct importation or adaptation.

As regards the adaptation of foreign terms, Table 12 illustrates the greater tendency of BP to adapt foreign borrowings, in contrast with EP which tends to replace them by vernacular terms. The percentage of adapted borrowings from English in the field of football increases dramatically in BP between 1950 and 1970. For the overall 21 onomasiological profiles, we find 23 adaptations and 19 loan translations in BP against 6 adaptations and 14 loan translations in EP.

Table 13 presents some examples of adaptation/translation of English loans in BP, and their replacement by vernacular expressions in EP. The loan *offside* is an exception: it is replaced by a vernacular expression in BP, and is translated in EP.

Finally, the proven high influence of foreign terms in BP on the football and clothing terms may be associated with two sociolinguistic phenomena which are taking place in Brazil. On the one hand, attitudes of linguistic purism are emerging (see section 2). On the other hand, attempts to implement them seem to be a complete failure, which is fully supported by the results presented in this section.

4.3. Other item-related features

We will now measure the proportions of three other item-related features: archaism, neologism and Brazilian terms. BP is the variety with the largest number of football terms that became obsolete between 1950 and 2000. The results leave no doubt at all (the 'archaism' feature identifies the terms that become obsolete in the time period considered): A'_{arch} (B50) = 21.7% and A'_{arch} (B00) = 0.2%, against A'_{arch} (P50) = 5.7% and A'_{arch} (P00) = 2.1%. If we include in this calculation all the foreign borrowings that become obsolete, asymmetry becomes even stronger: the percentage of A'_{arch} (B50) increases by 34.4%, while the percentage of A'_{arch} (P50) remains the same with 6%. To a certain extent, this result contributes to the convergence in the field of football. These changes, together with the changes observed earlier as regards the question of foreign borrowings are suggestive of greater flexibility from the Brazilian variety.

As to neologisms, it is of more interest to see their influence on the vocabulary related to clothing. It is not a surprise that new clothing terms or new uses of pre-existing terms are introduced in both varieties. These new terms or uses contribute to the global divergence between the two varieties, although they are not the only determinant factor. In fact, if we exclude recent items from the calculations, the results continue to indicate divergence, to a smaller extent and, yet, significance between 1970 and 2000: U' (P50,B50) 78.41% > U'(P70,B70) 69.62% > U' (P00,B00) 67.43%. The impact of recent concepts will be analyzed in section 4.5.

Finally, Table 14 reveals that the anticipated growing influence of BP on EP, particularly in the field of football, is not clearly confirmed. Two measurements of the Brazilian terms in the corpus of EP are given in this table: the percentages on the left include widely known Brazilian terms and the ones indicated on the right show all the Brazilian terms registered in reference dictionaries. As regards the percentage of Brazilian terms used in the European variety, the weighted measure varies from 0.8% to 2.3% in P50 and from 1.1% to 2% in P00. Examples of Brazilian terms which are introduced in EP are *bandeirinha* (ASSISTANT REFEREE), *atacante* (FORWARD), *falta* (FOUL), *plantel* (TEAM), among others. There are Brazilian terms, however, for which no occurrence was found at all in the corpus of EP: for instance, *goleiro* (GOALKEEPER), *arqueiro* (GOALKEEPER), *avante* (FORWARD), *escanteio* (CORNER), *impedimento* (OFFSIDE), *arco* ($GOAL_2$), *gol* ($GOAL_1$, $GOAL_2$), and *zagueiro* (BACK). For this reason, the feature of the Brazilian terms probably had little impact on the global convergence between the two varieties. On the other hand, the percentage of European Portuguese terms – as, for example, *guarda-redes* (GOALKEEPER), *baliza* ($GOAL_2$) or *defesa* (BACK) – in the corpus of BP is nil, and, therefore, confirms the hypothesis that the European variety has no influence whatsoever on the Brazilian variety.

Table 14. Brazilian terms in the corpus of European Portuguese for football

A'$_{Braz}$ (P50)	0,8%	2,3%
A'$_{Braz}$ (P70)	1,0%	3,4%
A'$_{Braz}$ (P00)	1,1%	2,0%

Brazilian terms constitute exogenous terms in the European variety. When EP adopts these terms, the proportion of exogenousness increases, as expected. Comparing Table 14 to Table 6 and 7, concerning endo-/exogenous

parameters in the vocabulary related to football, we can see that the low percentage of Brazilian terms in EP is reflected by the low exogenousness values in this variety in each period and by small changes of these values through time. Brazilian terms are not the only exogenousness factor in EP (or of endogenousness in BP). The sequential calculation in Table 7 shows that there is a stronger increase of exogenousness in BP than in EP. Therefore, the feature of the Brazilian terms does not seem to have influenced greatly the global convergent tendency between the two varieties.

4.4. The mutual impact of uniformity and item-related features

We will now examine the extent to which the main item-related features (that is, endo-/exogenousness and loanwords) influence the uniformity U values and, then, contribute to the global evolution. A first answer is given with the graphs of Figures 3 and 4. The results obtained for each variable are weighted measurements (ENGL>PORT identifies English terms that were adapted and LOAN represents all the loanwords in general). A negative correlation is always observed between U' and ENDO', and a negative correlation between U', EXO', ENGL' and LOAN', except in the field of football as regards the European variety (where the correlation is positive). A positive correlation is observed between U' and FREN' in the field of clothing. Therefore, we have the same pattern for both varieties in the field of clothing and the same one for both lexical fields, except for the European variety. This means that uniformity U decreases when ENDO, EXO, ENGL and LOAN increase, and uniformity U rises when these item-related features lose ground.

   The first two correlations do not come as a surprise: as we have seen in 4.1., the increase in uniformity corresponds to a decrease in endogenousness and exogenousness. There is an exception, though, in the vocabulary concerning football in EP. The negative correlation between U and ENG/LOAN is associated with the asymmetry between the two varieties in their accessibility to import and adapt loanwords, in general, and English loans, in particular (BP being more accessible than EP). The positive correlation between these same features in the football vocabulary in EP can also be associated with this asymmetry, because if more English loans exist in EP then the uniformity with BP is higher. Furthermore, the positive correlation between U and FREN in the clothing vocabulary in both varieties can be interpreted as a specific effect of the semantic field in question; both varie-

ties tend to behave in the same way with regard to fashion and clothing French terms.
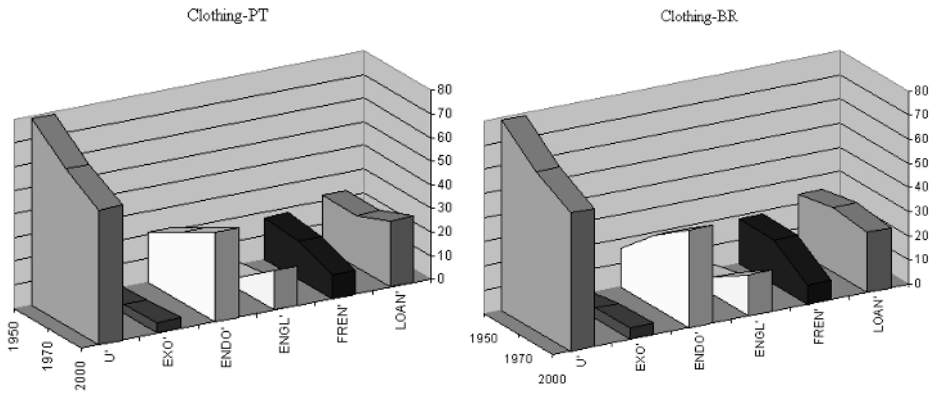
Figure 3. Correlation between uniformity and item-related features in the football corpus
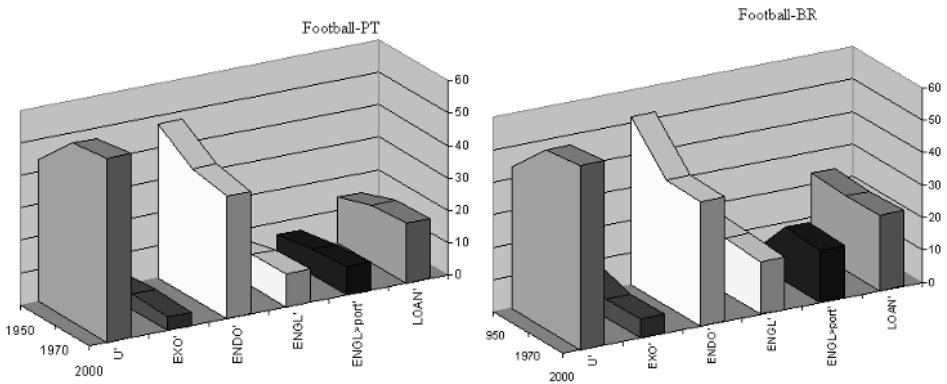
Figure 4. Correlation between uniformity and item-related features in the clothing corpus

Statistics give a more precise answer in Table 15 through the method of regression analysis.[4] The ENDO (endogenousness) parameter appears to be the most correlated one with uniformity U: it always provides significant

statistical results (or very close to significance) and about 99% of the evolution of U' can be described through the A'ENDO formula. Note that this correlation is more significant in BP than in EP. This strong correlation between endogenousness and uniformity derives from the actual mathematical definition of uniformity itself. The ENGL (English terms) and LOAN (all loanwords) parameters are also significant in the football vocabulary for the European variety. The other correlations do not pass the *p* value test. Nevertheless, the results obtained from linear regression have to be viewed carefully, given the rather low number of factors studied (5 parameters and 3 time periods only).

Table 15. Correlation between uniformity and item-related features

Football - Portugal	*p*	*r²*	Clothing - Portugal	*p*	*r²*
U' = -69,43 + 27,33 A'EXO	0,6389	28,87	U' = 139,7 - 19,03 A'EXO	0,2462	85,78
U' = 110,6 - 1,406 A'ENDO	0,0734	98,68	U' = 105,142 - 1,292 A'ENDO	0,0698	98,8
U' = 13,98 + 4,199 A'INGL	0,004	100	U' = 78,991 - 1,352 A'INGL	0,2916	80,45
U' = 14,36 + 4,94 A'INGL>PORT	0,1488	94,63	U' = 22,57 + 3,069 A'FRAN	0,2613	84,08
U' = 4,433 + 2,831 A'ESTR	0,0049	99,99	U' = 165,8 - 4,064 A'ESTR	0,4298	60,94
Football - Brazil	*p*	*r²*	Clothing - Brazil	*p*	*r²*
U' = 74,97 - 3,292 A'EXO	0,1254	96,17	U' = 93,84 -7,141 A'EXO	0,2355	86,93
U' = 91,41 - 0,9214 A'ENDO	0,0555	99,24	U' = 92,86 -0,8893 A'ENDO	0,0171	99,93
U' = 198,6 - 8,581 A'INGL	0,2617	84,03	U' = 84,23 - 1,766 A'INGL	0,237	86,77
U' = 41,07 + 0,9099 A'INGL>PORT	0,1079	97,15	U' = 34,68 + 2,197 A'FRAN	0,3808	68,28
U' = 765,8 - 30,77 A'ESTR	0,5594	40,72	U' = 462 -15,74 A'ESTR	0,6861	22,4

4.5. Concept-related features

So far we have analyzed various features associated to alternative terms within the onomasiological profiles of various concepts. We will now evaluate the impact of some concept-derived features on the global evolution trend. Three concept-related features will be looked at: semantic field, frequency and recent origin.

As regards semantic field, this feature has demonstrated that football and clothing terms indicate different results as far as convergence/divergence is concerned: evidence has shown restricted convergence in the semantic field of football and a clear divergence in the semantic field of clothing. Therefore, we have to analyze other semantic fields. However, both lexical fields indicate identical results as regards other aspects of lexical relationships between the two varieties of Portuguese.

As regards frequency, uniformity has seen to be higher with the more frequent concepts in both semantic fields: as we observed earlier, U' values are always higher (generally by 10%) than U values. This means that the more frequent concepts show better the global evolution trend (in the case of clothing, less frequent concepts also indicate divergence).

Finally, as regards recent concepts, three of them appear in the field of clothing after 1950: T-SHIRT, JEANS and LEGGINGS. The uniformity U of the first one is rather low because existing items were selected to name this new concept. The U value of JEANS and LEGGINGS is somewhat higher than the U average in 2000. Despite the limited empirical basis (only 3 concepts), it can be said that the divergent evolution in the vocabulary of clothing starts only after a certain period of adaptation.

5. Synchronic analysis: stratification results

Let's now analyze the question of stratification or current distance between the standard strata and the substandard strata (not strictly dialectal, but of an intermediate level). To this purpose, we will compare data in the field of football from the formal register of newspapers of the 1990s/2000s (P00 and B00) and data from the informal register of Internet chats ($P_{sub}00$ and $B_{sub}00$). In the case of clothing, we will compare the data of the more formal and national register of fashion magazines from the 1990s/2000s (P00 and B00) and data from the informal and more local register of the labels pictured from clothes shops ($P_{sub}00$ and $B_{sub}00$). The extension of the two sub-corpora (chats and labels) of the substandard strata and the respective number of observations were already mentioned in section 2. The results are as follows (on the left are the unweighted U and weighted U' uniformity percentages for EP; on the right are the unweighted U and weighted U' uniformity percentages for BP):

Football:

$U(P00,P_{sub}00)$	$\cong$	$U(B00,B_{sub}00)$
79.48%	$\cong$	82.18%
$U'(P00,P_{sub}00)$	$\cong$	$U'(B00,B_{sub}00)$
80.93%	$\cong$	78.76%

Clothing:

$U(P00,P_{sub}00)$	$>$	$U(B00,B_{sub}00)$
59.94%	$>$	51.71%
$U'(P00,P_{sub}00)$	$>$	$U'(B00,B_{sub}00)$
70.99%	$>$	57.94%

The hypothesis of a greater distance between the standard and the substandard strata in the Brazilian variety is confirmed in the sub-corpus of clothing but not in the sub-corpus of football.

In the case of football, there is quite a high uniformity result between the standard (newspapers) and the substandard (chats) in each variety, which may suggest lexical standardization of the Internet chats. Internet chat users tend to be conformist when they name concepts as popular as football. Furthermore, the informal language used in Internet chats tends to be determined by linguistic factors other than lexical factors.

We are now able to report on the level of uniformity between the two varieties at the level of the substandard strata and compare it with the uniformity at the level of the standard data in the same period. In the case of football, the uniformity value between the two varieties at the level of the substandard strata (59.21%) is as low as the uniformity value of the standard strata (56.76%, in Figure 1): formally, U' ($P_{sub}00$, $B_{sub}00$) 59.21% $\cong$ U' (P00, B00) 56.76%. In other words, the substantial distance between both varieties is verified both at the level of the standard strata and substandard strata. In the case of clothing, it has to be said that the distance between the two varieties is even wider at the substandard level (45.78%) than at standard level (57.11%, in Figure 2): U' ($P_{sub}00$, $B_{sub}00$) 45.78% $<$ U' (P00, B00) 57.11%. Although uniformity between the two varieties at the substandard level was not measured for periods prior to 1990/2000, the low value of uniformity at the substandard level reflects the clear divergent evolution between the two varieties.

6.   Conclusions and further research

This study of the lexical relationships between the European and Brazilian Portuguese varieties in the last 60 years allows us to draw a few conclusions. First, the results obtained for the fields of football and clothing differ with regard to the main issue of the present study and illustrate the complexity of the study of convergence and divergence between differing national varieties of a transcontinental language.

Second, the hypothesis of divergence is confirmed in the lexical field of clothing but not in the lexical field of football. Clothing terms are more representative of common vocabulary and, therefore, the results obtained for clothing are probably closer to the sociolinguistic reality. The slight convergence observed in the field of football between the 1950s and 1970s is probably the effect of globalization and standardization of the vocabulary of football. In fact, the global evolutionary relation between the two varieties and the influence of internal linguistic features on the evolution are clearer and more homogeneous in clothing terms than in football terms. As a whole, there are a lot of differences between the two varieties: the uniformity value is only 57% in both lexical fields at the moment.

Third, it seems that there is no specific orientation from one of the varieties towards the other. Both varieties diverge from each other in the vocabulary of clothing. The slight move of the Brazilian variety towards the European variety in the vocabulary of football results mainly from the adaptation of foreign borrowings that were massively introduced in their original form in BP in the 1950s. The influence of the Brazilian variety on the European variety, especially in the vocabulary of football, is not as clear as expected.

Fourth, the Brazilian variety has changed more than the European variety: is the greater changeability of the Brazilian variety the effect of greater external complexity and greater social variation or the effect of more recent standardization? Most probably all of them play a part. On the other hand, more changes occurred between the 1950s and the 1970s in both varieties and in both lexical fields. Fifth, it is proven that the influence of English and other foreign languages is stronger in the Brazilian variety: Brazilian Portuguese imports a larger number of loanwords and adapts and integrates them more easily than European Portuguese. Sixth, the clothing terms confirm the hypothesis of the synchronic stratificational asymmetry of the two varieties, especially the hypothesis that the actual distance between the standard and the substandard strata is higher in BP than in EP.

Finally, we hope to have shown the advantages of the cognitive perspective and of corpus-based and quantitative methodology (particularly the onomasiological perspective and profile-based methodology) for the study of the relationships between these two national varieties of Portuguese. We hope, in other words, to have contributed to the development of Portuguese cognitive sociolinguistics.

Of course, more research needs to be done into the issue of convergence and divergence between the two national varieties of the Portuguese language. As a present extension of this study, we intend to include (i) words from other lexical fields, such as health, and (ii) function words, namely prepositions. Prepositional profiles are limited to the same complements and syntagmatic context in order to satisfy the denotational synonymy condition. As a future extension to this study, we wish to include the field of grammar, to analyze non-lexical variables, namely syntactic and morphological ones. The aim is to study correlations between lexical and non-lexical variables and compare their impact on convergence/divergence and the stratification of the two national varieties of the Portuguese language. Given the attentional distinction between lexicon (more awareness) and syntax (less awareness), the hypothesis is that function words and syntactic constructions diverge more than content words.

We could not conclude this paper without pointing out briefly the main differences and similarities in the process of lexical convergence and divergence between the present study and that of Geeraerts, Grondelaers, and Speelman (1999) for Dutch. The main similarities are the higher significance of weighted measures and the more frequent onomasiological profiles, in the sense that they reflect the evolutionary trends in a better way than the less frequent profiles. Other similarities are the correlation between global evolutionary tendencies and internal linguistic features, the impact of foreign influence and the synchronic stratificational asymmetry. With regard to the differences between the studies, essentially two can be singled out. First, while the two national varieties of Dutch are clearly convergent, the two national varieties of Portuguese tend to be divergent. More importantly, the whole evolutionary situation (external and internal) is clearer and more homogeneous in Dutch than in Portuguese. Second, unlike the Dutch case where the Belgian variety clearly shows an exogenous orientation to conform to the Netherlandic variety, the Portuguese case does not exhibit any global unilateral orientation of one variety towards the other. In both languages, one variety exhibits more movement, diachronically speaking, than the other, but contrary to the Belgian variety, the movement of the

Brazilian variety does not appear to occur with its counterpart as reference point. Naturally, the differences in the lexical convergence/divergence process between Portuguese and Dutch reflect the geographical and soci-ocultural and historical differences of these pluricentric languages, which are clearly stronger and more complex in Portuguese than in Dutch.

Notes

1. This study is part of the "Convergence and Divergence in the Portuguese Lexicon" research project, funded by the Portuguese Foundation for Science and Technology (POCTI/LIN/48575/2002) in the framework of the POCTI of the European Communitarian Fund FEDER. I would like to thank three anonymous reviewers for their thorough and illuminating comments. I would also like to thank Stefan Grondelaers, Dirk Speelman, Dirk Geeraerts and Yves Peirsman for their methodological support and their useful comments and suggestions during the various stages of this project.
2. The two lists of terms presented in the appendix were based on Portuguese reference dictionaries, football glossaries, Feijó's studies (1994, 1998) on the language of football, Farias' (2003) fashion glossary and also on the content of the corpus which was built for this study. Loanwords that keep their original form are indicated in inverted commas.
3. Due to space limitations, the results of the total number of the terms studied (183 football terms and 264 clothing terms) cannot be presented here.
4. The *p* value indicates whether the relationship under analysis is statistically relevant: statistically, a relationship is significant only if *p* is lower than 0.05 (a value between 0.05 and 0.10 can be viewed as indicative of a trend). And the r^2 value indicates the percentage of data explained by a particular variable.

References

Bagno, Marcos
    2001   *Português ou Brasileiro? Um convite à pesquisa*. São Paulo: Parábola Editorial.
Belard, Francisco
    2001   Brasil e Portugal. *Jornal Expresso*, May 5 2001, Cartaz, 66.
Brandão, Sílvia Figueiredo, and Maria Antónia Mota (eds.)
    2003   *Análise Contrastiva de Variedades do Português. Primeiros Estudos*. Rio de Janeiro: In-Fólio.

Castilho, Ataliba Teixeira de

2005   Língua portuguesa e política linguística: O ponto de vista brasileiro. In *A Língua Portuguesa: Presente e Futuro* 193-221. Lisboa: Fundação Calouste Gulbenkian.

Castro, Ivo

1986   Um juízo sobre o novo Acordo Ortográfico. *ICALP* 5: 41-48.

Faraco, Carlos Alberto (ed.)

2004   *Estrangeirismos. Guerras em Torno da Língua*. São Paulo: Parábola Editorial.

Farias, Emilia Faria Peixoto

2003   *Glossário de Termos da Moda*. Fortaleza: Editora UFC/Edição Sebrae/CE.

Feijó, Luiz Cesar Saraiva

1994   *A Linguagem dos Esportes de Massa e a Gíria no Futebol*. Rio de Janeiro: UERJ-Tempo Brasileiro.

1998   *Brasil x Portugal: Um Derby Lingüístico*. Rio de Janeiro: Sociedade Brasileira de Língua e Literatura.

Geeraerts, Dirk

2001   On measuring lexical convergence. In *Linguagem e Cognição: A Perspectiva da Linguística Cognitiva*, Augusto Soares da Silva (ed.), 51-61. Braga: Associação Portuguesa de Linguística and Universidade Católica Portuguesa.

2005   Lectal variation and empirical data in Cognitive Linguistics. In *Cognitive Linguistics. Internal Dynamics and Interdisciplinary Interaction*, Francisco J. Ruiz de Mendoza Ibáñez, and M. Sandra Peña Cervel (eds.), 163-189. Berlin/New York: Mouton de Gruyter.

Geeraerts, Dirk, Stefan Grondelaers, and Peter Bakema

1994   *The Structure of Lexical Variation. Meaning, Naming, and Context*. Berlin/New York: Mouton de Gruyter.

Geeraerts, Dirk, Stefan Grondelaers, and Dirk Speelman

1999   *Convergentie en Divergentie in de Nederlandse Woordenschat*. Amsterdam: Meertens Instituut.

Kristiansen, Gitte, and René Dirven (eds.)

2008   *Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems*. Berlin/New York: Mouton de Gruyter.

Mattos e Silva, Rosa Virgínia

2004   *Ensaios para uma Sócio-História do Português Brasileiro*. São Paulo: Parábola Editorial.

Peres, João Andrade, and Mary Aizawa Kato (eds.)

2001   Colóquio Português Europeu / Português Brasileiro: Unidade e Diversidade na Passagem do Milénio. In *Actas do XVI Encontro Nacional da Associação Portuguesa de Linguística*, Clara Nunes

Correia, and Anabela Gonçalves (eds.), 537-769. Lisboa: Associação Portuguesa de Linguística.

2004   *Studies in the Comparative Semantics of European and Brazilian Portuguese*. Special issue of the *Journal of Portuguese Linguistics* 3.

Santos Diana

2009   Caminhos percorridos no mapa da portuguesificação: A Linguateca em perspectiva. *Linguamática* 1 (1): 25-59.

Santos, Diana, and Luís Sarmento

2003   O projecto AC/DC: acesso a corpora/disponibilização de corpora. In *Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística*, Amália Mendes, and Tiago Freitas (eds.), 705-717. Lisboa: Associação Portuguesa de Linguística.

Soares da Silva, Augusto

2005   Para o estudo das relações lexicais entre o Português Europeu e o Português do Brasil: Elementos de sociolexicologia cognitiva e quantitativa do Português. In *Actas do XX Encontro Nacional da Associação Portuguesa de Linguística*, Inês Duarte, and Isabel Leiria (eds.), 211-226. Lisboa: Associação Portuguesa de Linguística.

2008a   Integrando a variação social e métodos quantitativos na investigação sobre linguagem e cognição: para uma sociolinguística cognitiva do português europeu e brasileiro. *Revista de Estudos da Linguagem* 16 (1): 49-81.

2008b   O corpus CONDIV e o estudo da convergência e divergência entre variedades do português. In *Perspectivas sobre a Linguateca / Actas do Encontro Linguateca: 10 anos*, Luís Costa, Diana Santos, and Nuno Cardoso (eds.), 25-28. http://www.linguateca.pt/LivroL10/

Speelman, Dirk, Stefan Grondelaers, and Dirk Geeraerts

2003   Profile-based linguistic uniformity as a generic method for comparing language varieties. *Computers and the Humanities* 37: 317-337.

Appendix

Football profiles

BACK: "(full-)back", beque, bequeira, defensor, defesa, lateral, líbero, zagueiro
BALL: balão, bola, couro(inho), esfera, esférico, pelota
COACH: mister, professor, técnico, treinador
CORNER: canto, chute de canto, "corner", córner, escanteio, esquinado, pontapé de canto, tiro de canto

DRIBBLING: corte, drible(ing), engano, "feint", finta, firula, ginga, lesa, manobra enganadora, simulação

FORWARD: atacante, avançado, avante, dianteiro, "forward", ponta-de-lança

FOUL: carga, falta, "foul", golpe, infra(c)ção, obstru(c)ção, transgressão, violação (das regras)

FREE KICK: chute (in)direto, falta, "free(-kick)", livre (directo, indirecto), pontapé livre, tiro dire(c)to, tiro livre (direto, indireto)

$GOAL_1$: bola, "goal", gol, golo, ponto, tento

$GOAL_2$: arco, baliza, cidadela, "goal", gol(o), malhas, marco, meta, rede, redes, vala

GOALKEEPER: arqueiro, "goal-keeper", goleiro, golquíper, guarda-meta, guarda-rede, guarda-redes, guarda-vala, guarda-valas, guardião, "keeper", porteiro, quíper, vigia

MATCH: batalha, choque, combate, competição, confronto, desafio, disputa, duelo, embate, encontro, jogo, justa, luta, "match", partida, peleja, prélio, prova, pugna

MIDFIELDER: alfe, central, centro-campista, centro-médio, "half", interior, médio, meia, meio-campista, meio-campo, "midfield", trinco, volante

OFFSIDE: adiantamento, banheira, deslocação, fora-de-jogo, impedimento, "offside", posição irregular

PENALTY: castigo máximo, castigo-mor, falta máxima, grande penalidade, penalidade, penalidade máxima, penálti (pênalti, pénalti), "penalty"

REFEREE: apitador, árbitro, director da partida, juiz, juiz de campo, "ref(eree)", referi, refre

ASSISTANT REFEREE: árbitro auxiliar, árbitro assistente, auxiliar, 2º/3º/4º árbitro, bandeirinha, fiscal de linha, juiz de linha, "liner"

SHOT/KICK: chute, chuto, "kick(-off)", panázio, pelotada, pontapé, quique, "shoot", tiro

SHOT/PLAYING: jogada, lance

TEAM: conjunto, formação, eleven, equipa/e, esquadra, esquadrão, grupo, "match", onze, onzena, plantel, quadro, "team", time, turma

WINGER: ala, extremo, ponta, ponteiro

Clothing profiles

BLOUSE F: "blouse", blusa, blusinha, "bustier", camisa, camisa-body, camisão, camiseiro(inho), camiseta/e, (blusa) "chémisier", (blusa) chemisiê

CARDIGAN M/F: cardigã, "cardigan", casaco/casaquinho de malha (de lã, de tricô), "gilet", japona, malha, "twin-set"

COAT F: "blazer", blêizer, casaco, casaquinho/a, "manteau", mantô, paletó, "paletot"

COAT M: "blazer", blêizer, casaco, paletó, "paletot"

DRESS F: camiseiro, "chemisier", chemisiê, "shirt-dress", traje/o, veste, vestido(inho), vestido-camisa, vestido-camiseiro, vestido-camiseta, vestido-chemiser(ê), (vestido) cai-cai, (vestido) tomara-que-caia

JACKET M/F: casaca, casaco curto, jaleca, jaqueta, "jaquette", jaquetinha, véstia

JACKET (BLOUSON) M/F: "blazer", blêizer, blusão, "bluson", camurça, camurcine, camisa esporte, casaco de pele (de ganga, etc.), colete, parka

JEANS M/F: calça(s) de ganga, calça(s) em denim, calça(s) em jeans, ganga, jeans

JUMPER M/F: blusa, blusão, blusinha, "body", cachemir, camisa, camisa-de-meia, camiseta, camisinha, camisola, camisolinha, "canoutier", canoutiê, malha, malhinha, moleton, "pull", "pullover", pulôver, suéter, "sweat", "sweat shirt", "sweater"

LEGGINGS F: "fuseau(x)", fusô, "legging(s)"

OVERCOAT M/F: abafo, agasalho, balandrau, capote, casacão, casaco comprido, casaco de abafo/abafar, casaco de agasalho, casaco de/em pele, casaco-sobretudo, "duffle-coat", gabão, "gilet", "manteau", mantô, manto, overcoat, paletó, "pardessus", "pelerine", samarra, sobrecasaca, sobretudo, sobreveste, "trench (coat)"

RAINCOAT M/F: "ciré", "ciré-maxi", "anorak", canadiana, capa, capa de chuva, casaco impermeável, corta-vento, casaco-gabardina, gabardine/a, impermeável, kispo, parka

SHIRT M: blusão, camisa, camisa de gravata, camisa de manga curta, camisa desportiva, camisa esporte(iva), camisa jeans, camisa social, camiseta, camisete, "camisette"

SHORT JACKET F: bolero, carmona, casa(i)b(v)eque, casaco curto, casaquilha, colete, colete camiseiro , corpete, corpinho, garibáldi, "gilet", manguito, mini, minicasaco, roupinha, "shortie", vasquinha

SHORT JACKET M: casaco curto, colete, espartilho, gibão, "gilet", jaleca, jaleco, jaqueta, véstia

SHORT TROUSERS M/F: bermuda(s), calças-capri, calça(s) corsário, calça(s) curta(s), calças 3/4, calções, "cool pants" , corsários, "hot pants", "knikers", "pantacourt", "pedal pusher", "short(s)", "short cuts", "short shorts", shortinho, "slack(s)"

SKIRT F: kilt, maxi (máxi), maxissaia, micro-mini, micro-saia, míni (mini), mini-saia, minissaia, pareô, saia, saia-calça, saia-calção, saião, sainha, saiote

SUIT M: beca, completo, costume, fato, terno

SUIT/OUTFIT F: "complet", completo, conjunto, costume, duas-peças, "ensemble", fatinho, fato, saia-casaco, "tailleur", "toilette", toilete, vestido-casaco

TAILORED JACKET M/F: "black-tie", casaca, casaco cerimónia, fraque, "manteau", mantô, paletó, "paletot", "pelerine", "smo(c)king", sobrecasaca, "tuxedo"

TROUSERS M/F: calça, calças, pantalona

T-SHIRT M/F: camisa, camiseta/e, "camisette", camisola, licra, "singlet", "tee-shirt", "t-shirt"

Awesome insights into semantic variation

Justyna Robinson

Abstract

This paper demonstrates the benefits of employing a Cognitive Sociolinguistic approach in the investigation of lexical polysemy. The discussion is based on the usage-based analysis of the adjective awesome in a speech community. The results indicate that polysemy is far from a stable phenomenon, both at the conceptual level and at the sociolinguistic level. This work also suggests that combining cognitive and sociolinguistic analytical methods provides helpful insights into the flexibility of a polysemous category.

Keywords: semantic variation, semantic change, *awesome*, polysemy, apparent time

1. Introduction

The need for a socio-cognitive orientation in linguistic research has recently been advocated within Cognitive Linguistics (see Geeraerts 2005, Kristiansen and Dirven 2008). It is argued that both disciplines - Cognitive Linguistics and Sociolinguistics – share a common ground. With regard to the subject matter, both are interested in learning about the motivation for speakers' linguistic choices. From the point of view of methodology, both take a usage-based, empirical approach to the analysis of language data.

This chapter* constitutes a contribution to the discussion of the validity of Cognitive Sociolinguistics in language research, with a special focus on semasiological variation. I demonstrate that the socio-cognitive approach provides further insights into the flexibility of a polysemous category, especially in the context of diachronic changes of meaning.

The discussion revolves around the analysis of semantic variation of the adjective *awesome*. The analysis on the conceptual level is further complemented with information on the speakers' age, gender, and socio-economic

status. The results are explored within the appropriate cognitive and variationist paradigms.

2. Review

Sociolinguistics assumes that language variability mirrors social structure (Weinreich, Labov, and Herzog 1968). Seminal works in that area include Labov's Martha's Vineyard (1963) and New York studies (1966, 1972) and Trudgill's Norwich study (1974). Since then, structured variation has regularly been examined in the context of phonology and morpho-syntax (see summary in Chambers, Trudgill, and Schilling-Estes 2002). However, conclusions concerning the onomasiological and semasiological aspects of linguistic structure rarely constitute a central theme of sociolinguistic research.

Although studies of lexis were carried out in the context of word geography (e.g. Orton and Dieth 1962, Peters 1988, Upton and Widdowson 1999) and borrowings (Poplack, Sankoff, and Miller 1988), onomasiological variation and change in the "Labovian" sense has only recently been explored (Boberg 2004). Interest in socio-semasiology also surfaces in recent studies. Successful attempts to investigate meaning variation were carried out within functional paradigms (Hasan 1989, 1992, 2009) or discourse analysis frameworks (Cheshire 2007, Macaulay 2005, 2006, Stenström 2000, Tagliamonte and D'Arcy, 2004, Wong 2002, 2008). Their findings indicate that meaning construction relates to socio-demographic dimensions and that semantic change can be motivated by speakers' desire to index different stances of their identity. These studies suggest that further exploration of semantic variation is worthwhile.

Cognitive Linguistics has always recognized meaning as the most important aspect of linguistic structure (Geeraerts and Cuyckens 2007:14). It considers semantic structure to reflect flexibly speakers' perceptions and adaptation to their interaction with a physical and cultural reality. In this context, one would expect the variation of meaning to be regularly on the agenda of cognitive research. Indeed, lexical meaning variation has been studied within this framework, especially in the context of corpus research (e.g. Gries 2006, Divjak 2006, Gries and Divjak 2009, Beeching 2005). However, the relation between lexical variation and external sociolinguistic factors has been approached rather rarely. The seminal work in this area is Geeraerts, Grondelaers, and Bakema's (1994) study of onomasiological

variation in the context of dialect development. As far as semasiological variation is concerned, findings also indicate that the use of different concepts can be explained by a variety of interacting factors, dialect being one of the most significant ones.

From the theoretical perspective, both Sociolinguistics and Cognitive Linguistics agree on the variable nature of meaning, but in practice semantics has been explored to a different extent in both of these frameworks. In Sociolinguistics, there are few established methods that have been designed to deal with socio-semantics specifically. In Cognitive Semantics claims are often made at an abstract level and usage-based sociolinguistic approaches that consider individual speech differences are still relatively infrequent.

3. Scope and method of research

Sociolinguistics and Cognitive Linguistics have rarely used one another's legacies. Thus, one might wonder if "a socio-cognitive enterprise" is at all worth pursuing in language research. The present study addresses this question by looking into lexical polysemy. Two issues in particular are examined here:

· Can variationist sociolinguistics provide any insights into the conceptual structure of polysemy and cognitive processes involved in meaning variation and change?
· How can Cognitive Semantics be of use for Sociolinguistics?

In order to explore these issues, the usage of polysemous adjectives in present-day English is analyzed from a socio-cognitive perspective. In this work, I focus on instances of the adjective *awesome* only. I investigate if (socio-demographically) different people have different conceptualizations of the polysemous adjective *awesome*.

In order to address the above-mentioned issues a method of investigation that would yield data suitable for socio-cognitive analysis is needed. From the traditional sociolinguistic perspective, we would need data from a representative sample of a speech community controlled for chosen socio-demographic factors, e.g. age, gender, or social class. A usage-based study would require a design that elicits actual choices of language users.

Data were elicited in one-to-one interviews with members of the same speech community. 72 English speakers from South Yorkshire ranging in

age from 11 to 94 years old were interviewed. All of the volunteers were born locally and the majority of them lived in Sheffield. The sample of participants was equally representative of socio-demographic factors. Participants belong to four generations: up to 18, 19-30, 31-60, and over 60 years old. The divisions do not represent equal periods (in terms of a number of years), as they rather correspond to points of general lifestyle changes that are observable in the community. Each of the age groups contains the same number of speakers (18), controlled for gender and socio-economic background.

In order to establish the usage of meanings (their type and frequency) of polysemous adjectives for individual speakers the following structure of questions was employed (here eliciting the use of *awesome*):

> Q: Who or what is *awesome*?
> A: Grand Canyon.
> Q: Why is the Grand Canyon - *awesome*?
> A: Because it takes your breath away

This elicitation procedure asks for a referent described by the adjective in question (Who or what is [ADJ]?) and then asks for a clarification (Why is [REFERENT] [ADJ]?). This method is a suitable way of eliciting the natural usage of words as compared to asking direct questions about the meaning (i.e. 'What does *awesome* mean?'), which potentially elicits more conscious, formal, or standard variants. The answer to the question "Why is the Grand Canyon *awesome*?" provides participants' clarification on the sense in which an investigated polysemy was used. This information was particularly useful when the meaning of the adjective could not be identified on the basis of the referent only (e.g. "an *awesome* helper" turned out to be a 'horrible, nasty' helper, not a 'great' one).

Moreover, participants were instructed to answer each question with the first thing that came to mind. By following a similar structure of questions with each participant, the use of polysemous adjectives in the same context was elicited. As a consequence, comparable and thus quantifiable samples of the usage of individual meaning variants were obtained.

All together 15 adjectives were included in the interview: 8 adjectives with recently developed sense extensions, 7 controlling variables (polysemous adjectives without recently developed sense extensions and (broadly understood) monosemous adjectives).

In this paper, I report findings on one of the investigated variables, namely the adjective *awesome*. The structure of the polysemous category is

first explored and then analyzed in the context of cognitive and sociolinguistic paradigms.

4.   Case study: *awesome*

The analysis of the adjective *awesome* is based on a corpus of interview data consisting of 188 observations (each participant provided on average three instances of the investigated adjective). Responses involving the adjective *awesome* are manually categorized in terms of their usage into the following meaning groups. This is done mainly on the basis of matching the referents and associated explanations of usage to senses listed in dictionaries. The following meaning groups are identified:

- *Awesome* 'great': referential readings denoted referents of everyday close importance to a speaker. Examples of use: family, friends, pets, weekends, shopping.

- *Awesome* 'impressive': understood as 'overwhelming, breathtaking'; usually referring to something or someone exceptional, outstanding feats, breath-taking perceptual experiences. Examples of use: God, mountains, nature, Niagara Falls, sportsmen, opera singers, the Notre Dame Cathedral.

- *Awesome* 'terrible': 'frightening, weird, dreadful'. Examples of use: operations, a nasty helper, ugly person, war, bombings.

- *Awesome* "other": overlapping senses that could not be assigned reliably to any of the above groups. Examples of use: "Natural events, such as volcanoes and earthquakes, because they are beyond your control". The existence of overlapping uses accounts for the flexibility and non-discreteness of a semantic category. Meaning is fuzzy at edges and this property allows for the creation of conceptual extensions and links between senses which in usage-based research should surface as overlapping readings. The *awesome* "other" category also includes "reported" uses of the adjective *awesome*, i.e. cases when a participant states that s/he knows of a meaning but denies using it.

4.1. Cognitive semantics analysis of results

The usage of *awesome* in the investigated speech community is illustrated in Figure 1.

Awesome
'great
43%

Awesome
'impressive'
36%

Awesome
"other
9%

Awesome
'terrible'
12%
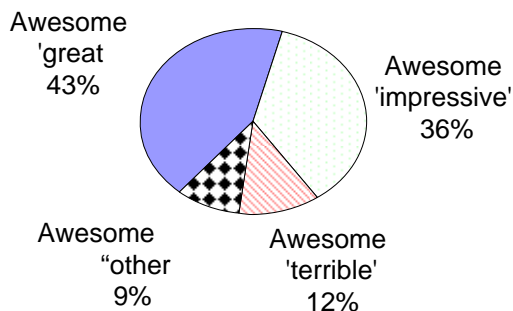
Figure 1. The use of the adjective *awesome*

The basic quantification of responses indicates that the most frequently used sense variant of this adjective is *awesome* 'great' (43% of responses) followed by *awesome* 'impressive (36% of responses). *Awesome* 'terrible', which occurs in only 12% of responses, seems to be a far less salient sense of the polysemous category in the investigated speech community. The remaining 9% of responses are ambiguous, overlapping, and reported cases grouped together in the category *awesome* "other".

Initial results confirm cognitive-semantic assumptions about the nature of meaning. The adjective *awesome* is a complex polysemous category consisting of a cluster of overlapping senses, with some of them more salient (such as *awesome* 'great') and others more peripheral (*awesome* 'terrible').

However, this purely conceptual analysis of responses is still not fully comprehensive. First, it is only satisfactory when one assumes the lack of dialect internal variation, that is, when we are happy with the statement that the language of a whole community is homogenous. Moreover, it is hard to read any patterns of change or tendencies in the development of the polysemous category (unless you take the most salient meaning as the "leading meaning"- but then how to interpret the chart, considering dictionary evidence suggesting that *awesome* 'great' is merely peripheral, slang?). In this context additional analytical steps are taken by adding a sociolinguistic element to the present data.

4.2. Socio-cognitive semantics analysis: *awesome*

In order to add a sociolinguistic facet to the analysis, participants' res-
ponses are analyzed in the context of age, gender, and socio-economic sta-
tus variables. Most of the discussion revolves around the category of age
since a number of variationist studies have shown its significance in struc-
turing and explaining linguistic variation (see e.g. Labov 1994, Eckert
1997, Romaine 1984). In addition, the metalinguistic comments of partici-
pants indicate that age might be an important factor in accounting for the
observed semantic variation. The effect of gender and socio-economic fac-
tors on the variation of the adjective *awesome* are discussed later in the
chapter.

   Figure 2 illustrates the usage of the conceptual category *awesome* plot-
ted as a function of age. Here the category "other" is split to represent indi-
vidual reported and overlapping senses. In comparison to the polysemy
structure illustrated in Figure 1, Figure 2 presents a far more dynamic pic-
ture of the usage of the adjective *awesome*. Each bar on the chart indicates
that the distribution of individual meanings is different for each age group.
For instance, *awesome* 'great' is the most salient reading for speakers that
are up to 18 years old, and the least frequent one for those over 60 years
old. This finding enhances our understanding of polysemy as a flexible
phenomenon. One can see that the flexibility of the same polysemous cate-
gory that was shown at the community level (Figure 1) also emerges for
individual age groups (Figure 2). However, the distribution of senses in
Figure 2 differs from the one in Figure 1. The structure of each bar in Fig-
ure 2 indicates that individual conceptualizations of the adjective *awesome*
are significantly different for different speakers within the same communi-
ty. (Kruskal-Wallis tests were carried out to assess whether the differences
in the mean usage of different senses of the adjective *awesome* in different
age groups are statistically significant. The results are as follows: *awesome*
'great' $p<.001$, *awesome* 'impressive' $p=.004$, *awesome* 'terrible' $p<.001$.)

   This observation clearly supports a two-fold aspect of polysemous flex-
ibility: (1) not every meaning is equally representative within a polysemous
category; (2) not every meaning is equally salient for every speaker, even
within the same speech community. The second remark is especially impor-
tant in the context of the value of socio-cognitive research. It extends the
findings of the prototypical nature of meaning (Rosch 1975, Geeraerts
1989) by demonstrating experimentally the great extent to which the salient
representations differ for individual speakers. Moreover, since each speaker

has different experiences of the world and an individually shaped understanding of a concept, the methods for the investigation of experiential meaning need to go beyond abstract conceptual approaches, towards usage-based ones. This analysis shows that employing sociolinguistic methods of research helps to manage (qualify and quantify) the individuality of linguistic observations.
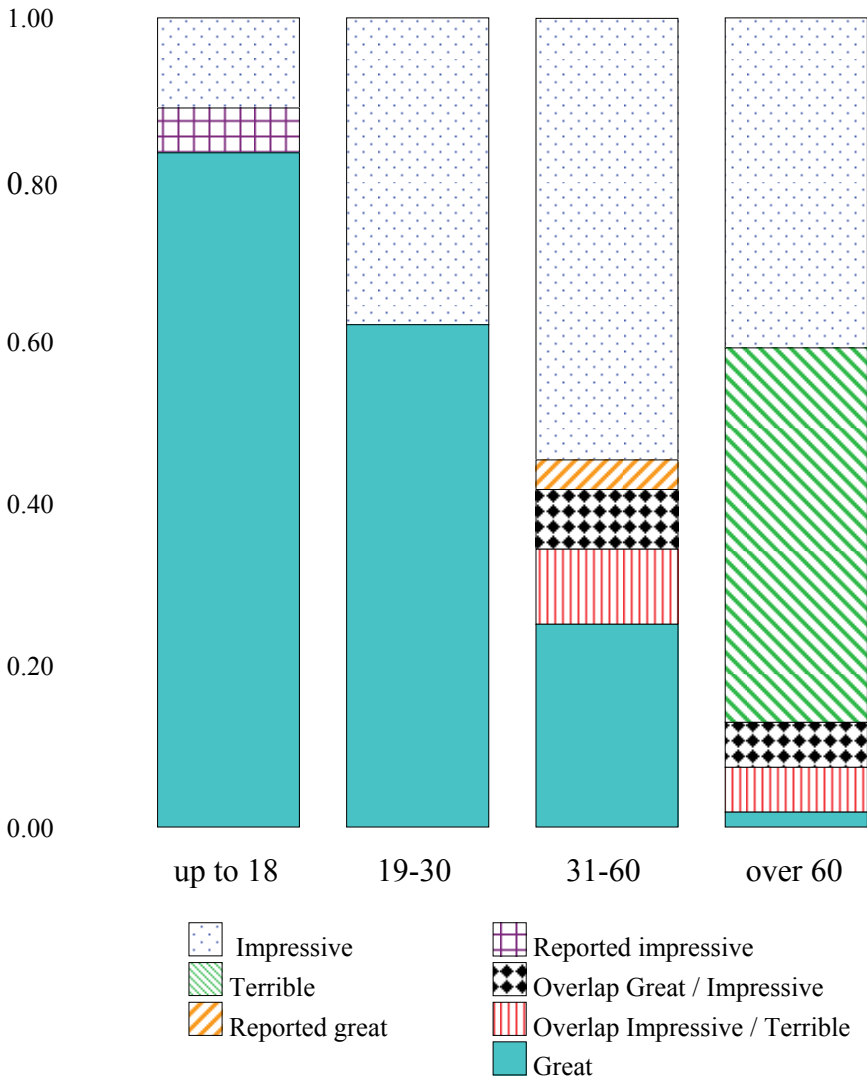
Figure 2. Age-related variation of the usage of the adjective *awesome*

Another observation of the socio-cognitive analysis of the adjective *awesome* is that the core meaning for some speakers is completely different from that of the community as a whole. The most salient sense at the community level (Figure 1) and for speakers up to 30 years old is *awesome* 'great'. But for speakers between 31 and 60 years old *awesome* 'impressive', and for those over 60 years old *awesome* 'terrible' are the most frequent uses of this polysemous adjective.

Thus, for half of participants the most salient meaning is different from what it is for a whole community. Why would this be the case? Especially, why would the core meaning of *awesome* be so different for speakers over 30 as compared to younger speakers? Of course, as was argued earlier, different speakers would have different individual conceptualizations. But how can we explain the fact that speakers within each generation share similar conceptualizations, which are yet significantly different between generations? This question is addressed in the context of the variationist paradigm, in which we take a closer look at the significance of socio-demographic categories.

4.3. Analysis of external factors in semantic variation

By mapping the conceptual structure (Figure 1) onto the demographic one, a clear pattern emerges (Figure 2): the use of particular senses is non-randomly related to the age of speakers. But to what extent does age actually explain the observed variation? Non-parametric statistics (Kruskal-Wallis tests) inform us only of the significant changes in mean usage between age groups. Also, is the speaker's age the only external factor that accounts for the use of polysemous *awesome*? A number of studies indicate that gender and the socio-economic status of speakers may also account for linguistic variation (see review of relevant studies in Coates 2004, Eckert 1998, or Kerswill 2006). Could such factors also account for the semantic variation of *awesome*? This section explores several socio-demographic variables in order to assess their significance in explaining the observed variation of the polysemous *awesome*. Speakers' age, gender, education, occupation, and their place of residence are taken into consideration.

In order to address the above-mentioned issues, a multifactorial statistical model needs to be employed, i.e. a model that considers several external factors simultaneously and measures their effect on the use of the different senses of *awesome*. In addition, the analysis has to be suitable for

testing dependent variables which are binary (e.g. use vs. non-use of *awe-some* 'great'). Finally, the appropriate statistical approach needs to allow us to check for nuisance variables which may confound the results. Socio-demographic factors may constitute such cases, e.g. education and occupation may be correlated, as people that are more educated are likely to have better jobs.

Table 1. Socio-demographic variables (independent variables) used in the logistic regression analyses to investigate their association with the use of different senses of the adjective *awesome* (dependent variables)

Independent variables	Coded as	Categories				
Age group	1-4	Up to 18	19-30	31-60	Over 60	
Gender	1-2	Male	Female			
NSEC	1-3	Higher	Medium	Lower		
Education	1-5	Prior to the age of 16	Secondary school	College/ 6th form	University	Currently a student
Neighborhood	1-3	Lower property prices	Middle property prices	Highest property prices		

Logistic regression analysis is the appropriate procedure to fulfill these requirements. Logistic regression is a mathematical modeling approach that can be used to test hypotheses about the relationship of several independent variables to a dichotomous dependent variable (see Hosmer and Lemeshow 1989, Kleinbaum 1994, Tabachnick and Fidell 2001 for an introduction to logistic regression). Usually it is used to predict a particular binary outcome (event or non-event) from a set of independent variables. For instance, you may want to use logistic modeling to assess if winning or losing a game of bridge can be predicted from the gender and years of experience of the players. Logistic regression also provides information on variation (the percentage to which an independent variable is explained by the dependent ones) and is used to determine the importance of the independent variables.

In this work, logistic regression is employed to assess the overall effect of socio-demographic factors on the use of the individual senses of the adjective *awesome*.

The following independent variables are considered in the logistic model: age group, gender, education, National Statistics Socio-Economic Classification (NSEC) score, which is based on occupation, and a neighborhood variable, which is based on property values in areas defined by the postcode of a participant's residence. For the summary of the coding of independent variables see Table 1.

Seventy-two cases with missing values were included in the analysis; equal numbers of males and females (36), and equal numbers of members of each age group (18). The factors age group and gender were entered into the model. Further factors were eliminated using a backward stepwise method. This was done because the prime interest was in age and gender. The other variables were allowed to stay if they significantly altered the model, as they were possibly confounding variables. The final model is reported.

4.3.1. Awesome 'great'

Logistic regression analysis is carried out to assess the overall effect of socio-demographic factors (independent variables) on the use of awesome *'great'* (dependent variable).

The summary of the results of the logistic regression analysis of the use of *awesome* 'great' is presented in Table 2. The final model reported includes the variables which best account for the observed variation. Insignificant variables are excluded from the model. The table shows the coefficients of regression Beta (B), their standard errors, the Wald Chi-Square statistic, associated p-values, and odds ratio (Exp(B)). The resulting fitted model indicates which independent variables are included in the final logistic model. It also informs us about significant changes in regression coefficients (B) between predictors. B determines the direction of the relationship between a given predictor and the dependent variable (the use of *awesome* 'great'). If B is positive, the odds of the use of awesome 'great' are increased; when B is negative, the odds are decreased; when B is 0, the odds stay unchanged.

Table 2. Logistic regression model for *awesome* 'great'

	Beta	S.E.	Wald	df	p	Exp(B)
AgeGroup			20.018	3	.000	
AgeGroup(1)[a]	.867	.962	.812	1	.368	2.379
AgeGroup(2)[b]	1.861	.792	5.514	1	.019	6.427
AgeGroup(3)[c]	2.505	1.162	4.650	1	.031	12.241
Gender(1)	-1.260	.689	3.343	1	.068	.284
Constant	.023	.376	.004	1	.950	1.024

[a]: indicator variable representing change between age group (19-30) in relation to age group (up to 18)
[b]: indicator variable representing change between age group (31-60) in relation to age group (19-30)
[c]: indicator variable representing change between age group (over 60) in relation to age group (31-60)

Main finding. According to the model, the age group contributes significantly to the model for the speakers' use of *awesome* 'great' (p<.001).

Age group. The most significant differences of use exist between age groups (31-60) and (19-30) (B=1.861, p=.019), and also between age groups (over 60) and (31-60) (B=2.505, p=.031). These results indicate that the two youngest age groups speak most similarly to each other. The use of *awesome* 'great' decreases significantly for speakers of age (30 -60) and then for those over 60 years old.

Gender. The model also includes gender. Females are more likely to use *awesome* 'great' at a marginally significant level (p=.068).

In the logistic regression analysis, the predictive and explanatory power of the fitted model needs to be assessed. In order to validate the predicted probabilities, the c-statistic is used (see Peng, Lee, and Ingersoll 2002:6). The c-statistic compares the proportion of observed cases to the probability of the occurrence of *awesome* 'great' that was initially predicted. In the case of *awesome* 'great', the fitted model (one that includes socio-demographic variables) achieves a success rate of 84.7%, which is an improvement over the intercept model, i.e. a model that does not include any of the socio-demographic variables to account for the observed variation, but includes a constant term only (52.8%).

The explanatory power of the calculated model refers to how effectively it fits the actual data for estimating the outcome variable (Moss, Wellman, and Cotsonis 2003: 925). This could be assessed by a number of "goodness-of-fit" measures. -2 Log Likelihood (-2LL) indicates the overall fit of the model. It reflects the significance of the unexplained variance in the model. Its lowering values indicate improvement of a model fit (increasing the likelihood of the observed results). R-square measurements (Cox and Snell, Nagelkerke tests) indicate how much variation the model actually explains. Sometimes these measures may yield different results (for further discussion see Field 2005: 239-240). The Hosmer and Lemeshow test is another measure that is considered by some researchers to be a more accurate measure for assessing the goodness-of-fit of the model (Peng, Lee, and Ingersoll 2002:6). It says how closely the observed and predicted probabilities match; insignificant results of the Hosmer and Lemeshow test signify a model that fits the data well.

In the case of *awesome* 'great' lowering -2LL (59.742) and an insignificant result of the Hosmer-Lemeshow test indicate that the model fits the data well and is more adequate for explaining this variation than models which do not consider socio-demographic factors. R-square measurements (Cox and Snell= .425, Nagelkerke =.567) indicate that the variation in the outcome variable is explained by the logistic regression model moderately well.

Logistic regression analysis evidences that the use of *awesome* 'great' can be satisfactorily modeled from the age and gender of speakers, although age has a more significant overall effect on the use of the given variable than gender.

4.3.2. Awesome 'impressive'

Logistic regression analysis is carried out to assess the overall effect of socio-demographic factors (independent variables) on the use of *awesome* 'impressive' (dependent variable). The summary of the results for the logistic regression is presented in Table 3. As to the overall assessment and validation of the fitted model, the c-statistics compare the proportion of observed instances to the predicted probabilities of the use and non-use of *awesome* 'impressive'. In the case of *awesome* 'impressive', the fitted model achieves a success rate of 79.2%, which is an improvement over the

entry model (a model that includes age group and gender only) and inter-cept models (69.4% and 50.0% respectively).

Table 3. Logistic regression model for *awesome* 'impressive'

	Beta	S.E.	Wald	df	p	Exp(B)
AgeGroup			11.977	3	.007	
AgeGroup(1)[a]	-2.572	.928	7.679	1	.006	.076
AgeGroup(2)[b]	-.646	.810	.636	1	.425	.524
AgeGroup(3)[c]	1.717	.868	3.919	1	.048	5.570
Gender(1)	1.262	.609	4.291	1	.038	3.534
Neighborhood	1.172	.382	9.404	1	.002	3.228
Constant	-2.318	.793	8.541	1	.003	.098

[a]: indicator variable representing change between age group (19-30) in relation to age group (up to 18)
[b]: indicator variable representing change between age group (31-60) in relation to age group (19-30)
[c]: indicator variable representing change between age group (over 60) in relation to age group (31-60)

Main finding. According to the model, age group and gender contribute signifi-cantly to the model for the speakers' use of *awesome* 'impressive'.

Age group. The most significant 'jumps' in B-coefficients exist between age groups (up to 18) and (19-30) (B=-2.572, p=.006) and age groups (over 60) and (31-60) (B=1.717, p=.048). These results indicate that speakers over 19 are most likely to use *awesome* 'impressive' as compared to younger participants.

Gender. The significant difference in Beta values exist between males and fe-males' use of *awesome* 'impressive' (p=.038, B=1.262). Males are more likely to use this sense.

Neighborhood. The fitted model also indicates that the inclusion of the neighbor-hood significantly alters the model (p=.002). However, this variable may be con-founding: the older the person, the richer.

The explanatory power of the fitted model is assessed by comparing it with the intercept and entry models. Increasing R-square measurements (Cox and Snell= .335, Nagelkerke =.447) suggest that the fitted model accounts for around 40% of the variation, which is a moderate outcome. Lowering (still not very low) -2LL (70.412) and insignificant results on the Hosmer and Lemeshow test indicate that the final model fits the data well and reduces unexplained variation in comparison to the entry model and the intercept model.

Logistic regression analysis evidences that *awesome* 'impressive' can be satisfactorily modeled from the participants' age, gender, and potentially the area where they live (neighborhood may be a confounding variable).

4.3.3. Awesome 'terrible'

A logistic regression analysis of *awesome* 'terrible' yields an unstable solution, so it is not possible to make predictions regarding the overall effect of external factors on the use of this meaning.

4.4. Discussion of results

4.4.1. Age related variation of awesome

Logistic regression modeling indicated that the overall effect of the age of participants on the use of *awesome* is significant. The use of *awesome* 'great' decreases with increasing age of participants; the use of *awesome* 'impressive' is lowest for the youngest generation. Although a stable logistic model could not be established for *awesome* 'terrible' we may at least fall back on the results of non-parametric tests, which show that this variable is used significantly more frequently by speakers over 60 years old (p<.001).

The significance of age in the observed variation may be interpreted within the apparent time construct as indicative of semantic change in progress.

The apparent time hypothesis indicates that linguistic differences among different age groups or generations reflect actual diachronic developments in language. In other words, linguistic trends observed in synchrony could

actually indicate ongoing changes. A central thought of this approach refers to the uniformitarian principle, which asserts that:

> "If there are relatively constant, day-to-day effects of social interaction upon grammar and phonology, (…) these influences continue to operate in the same way that they have in the past" (Labov 1972: 275)

Labov's (1963) analysis of variation and change on Martha's Vineyard is the seminal work in this area. He argues that the rise in the use of centralized variants of (ay) and (aw) in apparent time reflects actual diachronic changes in the use of the features on Martha's Vineyard.

The model confirmed predictions concerning past trends as evidenced in data collected for the Linguistic Atlas of New England (Kurath 1941) more than 20 years earlier (Labov 1963: 275-276). Real-time evidence (Pope 2002) corroborated that Labov's anticipated pattern of centralization of (aw) and (ay) continued in the way that the 1963 survey indicated. The apparent time construct has been shown to be a successful way of investigating language variation and change in different speech communities (see the summary of relevant studies in Bailey 2002, Sankoff 2006, or Tagliamonte and D'Arcy 2009).

Based on the apparent time construct, observed generational differences in the usage of different sense variants of the adjective *awesome* may be interpreted as semantic change in progress. One could argue that *awesome* 'great' is developing as the core meaning of the polysemous category. *Awesome* 'terrible' is likely to soon be considered archaic since it is only active for the oldest members of that speech community.

Available real time evidence confirms the apparent time prediction. For instance, the Oxford English Dictionary Online indicates that *awesome* 'awful, dreadful, terrible' was in use since the 16th century, much earlier than *awesome* 'impressive' and 'great'. First quotations of *awesome* 'overwhelming, impressive, breath-taking' in the OED Online date back to the 1960s and *awesome* 'good, great' does not seem to be recorded in British English until the early 1980s, which is also confirmed by data from the Lancaster-Oslo/Bergen Corpus of British English (1961), and the British National Corpus.

The present study only hypothesizes change in progress of the adjective *awesome* and the observed trends can only be verified in 40 – 50 years time. However, apparent time studies so far show that locating a gradient age distribution in a community under study virtually ensures the existence of change, whether or not age grading is also involved (Sankoff 2006: 111).

4.4.2. *Effects of Gender and Neighborhood*

The fitted models for polysemous *awesome* also include gender and neighborhood alongside the age group factor. The use of *awesome* 'great' can be modeled from the speech of females and the use of *awesome* 'impressive' can be modeled from the speech of males and from the type of neighborhood (measured by property value) speakers live in.

Considering historical information on the emergence of the relevant senses, one may interpret these findings as follows: males use the more standard meaning (*awesome* 'impressive'), females lead the use (change) of the newer sense extension (*awesome* 'great'). Also, since the area speakers live in could be interpreted as one of the indices of socio-economic status (Roberts 2001), one could suggest that middle class people would exhibit more standard usage as well. However, the interpretation of the effect of the neighborhood variable should be cautious, since this variable may be correlated with age: the older you are the richer you become.

These results fit well with findings from other variationist studies (see summary in Kerswill 2006). Women, especially middle-aged and located in the middle of the socio-economic continuum lead linguistic change. A more standardized speech is often characteristic of the upper classes.

4.5. Polysemy: a dynamic picture

In the context of the apparent time hypothesis it becomes apparent that Figure 2 does not only represent variation in the polysemous readings, but a dynamic picture of semantic change in progress. One can already notice the potential benefits of employing a socio-cognitive method; in other words, the benefits of mapping individual conceptualizations of a polysemous category onto a variationist context.

The fact that the change in progress was observed helps to shed light on the differences between individual, generational, and community representations of the core meanings of *awesome*. The representation of *awesome* at the community level (Figure 1) does not correspond to what is observed at the generational level (Figure 2) because each generation actually represents a different stage in the diachronic development of this category. Speakers' conceptualizations relate socio-culturally to the language of a particular period, which to a large extent appears to be "frozen" in their age. For instance, older speakers are more likely to reflect the language of

their generation which would basically involve forms that they learnt when they were younger, and subsequent modifications. Similarly, younger generations use a language that is an outcome of a usage "competition" between the forms that they learn from their parents and teachers, and new forms they negotiate in their peer groups. These different conceptualizations emerge at the socio-cognitive level as a variation of uses non-randomly distributed across the age axis.

Another observation concerns the nature of semantic change itself. Fine-grained variationist investigation provides evidence for the existence of conceptual links between successive senses (cf. Geeraerts 1997, Nerlich and Clarke 1992). In the pattern of change derived from the apparent time construct (*awesome* 'terrible'> 'impressive'> 'great') *awesome* 'terrible' and *awesome* 'impressive' seem to be connected by an element of intense emotion, and *awesome* 'impressive' and *awesome* 'great' are similar in the context of positive emotion. Conceptual links between senses can also be found in examples grouped in the category *awesome* "other". This category includes overlapping senses, which can (but do not have to) be the first signals of a particular category being used in a novel way. What is interesting here is that the examples represent an overlapping of senses that are diachronically successive (e.g. *awesome* 'terrible' and *awesome* 'impressive', but never *awesome* 'terrible' and *awesome* 'great'). Moreover, the overlapping readings do not occur where one would not expect them to occur. For example, overlap involving *awesome* 'terrible' (the oldest sense) does not occur in the speech of the youngest generation but in generations which use this sense (over 60) and in the neighboring younger generation (31-60) (see Figure 2).

Socio-cognitive analysis also helps shed light on the circumstances surrounding the emergence of new meanings. This approach permits us to identify speakers who are likely to be responsible for innovating and propagating novel uses of a word, but also those who most resist innovation. In the case of *awesome*, one can see that speakers of the age 31-60 are the first to use *awesome* 'great', and those over 60 are the last to still be using *awesome* 'terrible'. Logistic regression analysis showed that the innovative meaning can be modeled from the speech of women, and the more standard meaning can be modeled from the speech of men and potentially upper classes. These general observations open the door to a potentially far more insightful analysis of the circumstances in which a conceptual category successfully enters a social construct. One could analyze the positioning of innovators in the socio-economic structure of a given community and in-

vestigate the social practices and networks they are a part of. As a consequence, such an approach could potentially get us a step closer to showing how a new meaning gets experientially grounded.

The socio-cognitive analysis of *awesome* indicates that fluctuations in the usage of a semantic category systematically relate to the socio-demographic characteristics of a speech community. This observation positively contributes to the discussion on the validity of a socio-cognitive orientation in linguistic research.

5.   Summary and conclusions

This work explores and discusses the benefits of a socio-cognitive orientation in linguistic research by carrying out a usage-based investigation of the polysemous adjective *awesome*. Although the initial analysis revealed the complexity of the polysemous category, it was not until a socio-demographic facet was imposed onto our data that the fine-grained aspects of polysemous flexibility emerged.

The first striking observation was the multiplicity and complexity of conceptualizations at different levels of analysis of the speech community. The core meaning of the investigated polysemy varied significantly for individual speakers and individual generations even within the same, comparatively small speech community.

Also, we have found that this "conceptual mess" can be better organized and interpreted within a variationist framework. A complex polysemy structure characteristic for each participant and each generation started to form a regular pattern when it was mapped on the demographic structure of a whole speech community. The apparent time hypothesis, supported with adequate statistics, indicated that this pattern, as a function of linguistic usage and the age of speakers, in fact indicates semantic change in progress. Thus, a socio-cognitive approach enabled us to trace polysemous flexibility step-by-step, or should we say, speaker-by-speaker, until seemingly "meaningless" individual variations actually showed a real development of meaning in time.

This finding points to the importance of looking into individual speakers, and the way their conceptualizations and perceptions are grounded in the socio-demographic and cultural reality. Of course, this line of thinking is not novel for Cognitive Linguistics which has argued for the experiential nature of language for a considerable time. Unfortunately, there have been

few studies that put this perspective into practice and analyze the language and environment of individual speakers (cf. Dąbrowska and Street 2006). One of the reasons for this situation might relate to the fact that Cognitive-Linguistic methods of investigation are not well-suited to tackle the social aspects of language and Variationist Sociolinguistics struggles to capture the complexity of meanings with the available tools. Therefore, combing sociolinguistic and cognitive approaches fills these gaps by potentially bringing beneficial solutions to language research at a methodological level.

Finally, this study also indicates the necessity of including external factors in exploring conceptual variation. The analysis of the adjective *awesome* shows that semasiological variation is not free but regularly structured in terms of the age, gender, and social class of speakers. This finding contributes to observations of other scholars who demonstrate that the entrenchment of conceptual categories can be explained in relation to external factors.

The investigation of polysemous *awesome* presented above contributes to the discussion of the validity of an emerging field of Cognitive Sociolinguistic. The application of this interdisciplinary perspective appears to be beneficial at a methodological and an analytical level. By employing variationist methods, the multiplicity of individual conceptualizations is successfully elicited, organized, and quantified. Consequently, the analysis within a Cognitive Sociolinguistic paradigm achieves a better explanatory power when accounting for complex phenomena that are conceptually, pragmatically, and socio-culturally grounded. This orientation in research seems to be a natural step forward in order to account for the multiple functions and the flexibility of meaning.

Note

References

Bailey, Guy
    2002   Real and apparent time. In *The Handbook of Language Variation and Change,* Jack K. Chambers, Peter Trudgill, and Natalie Schilling-Estes (eds.), 312- 332. Oxford: Blackwell.

Beeching, Kate
    2005   Politeness-induced semantic change. The case of quand même. *Language Variation and Change* 17: 155-180.

Boberg, Charles
    2004   Real and apparent time in language change: late adoption of changes in Montreal English. *American Speech* 79 (3): 250- 269.

Chambers, Jack. K., Peter Trudgill, and Natalie Schilling-Estes (eds.)
    2002   *The Handbook of Language Variation and Change*. Oxford: Blackwell.

Cheshire, Jenny
    2007   Discourse variation, grammaticalisation and stuff like that. *Journal of Sociolinguistics* 11 (2): 155–193.

Coates, Jennifer (ed.)
    1998   *Language and Gender: A Reader*. Oxford: Blackwell.
    2004   *Women, Men and Language*. Third Edition. London: Routledge.

Coulmas, Florian (ed.)
    1997   *The Handbook of Sociolinguistics*. Oxford: Blackwell.

Dąbrowska, Ewa, and James Street
    2006   Individual differences in language attainment: comprehension of passive sentences by native and non-native English speakers. *Language Sciences* 28 (6): 604-615.

Divjak, Dagmar
    2006   Ways of intending: delineating and structuring near synonyms. In *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*, Stefan Th. Gries, and Anatol Stefanowitsch (eds.), 19-56. Berlin/New York: Mouton de Gruyter.

Eckert, Penelope
    1997   Age as a sociolinguistic variable. In *The Handbook of Sociolinguistics*, Florian Coulmas (ed.), 151-167. Oxford: Blackwell.
    1998   Gender and sociolinguistic variation. In *Language and Gender: A Reader*, Jennifer Coates (ed.), 64-75. Oxford: Blackwell.

Field, Andy
    2005   *Discovering Statistics using SPSS*. London: SAGE Publications.

Fisiak, Jacek (ed.)
    1988   *Historical Dialectology. Regional and Social.* Trends in Linguistics. Studies and Monographs 37. Berlin/New York: Mouton de Gruyter.

Geeraerts, Dirk
    1989   Introduction: Prospects and problems of prototype theory. *Linguistics* 27: 587-612.
    1993   Cognitive linguistics and the history of philosophical epistemology. In *Conceptualisations and Mental Processing in Language*, Richard Geiger, and Brygida Rudzka-Ostyn (eds.), 53-79. Berlin/New York: Mouton de Gruyter.
    1997   *Diachronic Prototype Semantics. A Contribution to Historical Lexicology*. Oxford: Clarendon Press.
    2005   Lectal variation and empirical data in Cognitive Linguistics. In *Cognitive Linguistics: Internal Dynamics and Interdisciplinary Interaction*, Francisco José Ruiz de Mendoza Ibáñez, and M. Sandra Peña Cervel, 163-189. Berlin/New York: Mouton de Gruyter.
Geeraerts, Dirk, Stefan Grondelaers, and Peter Bakema
    1994   *The Structure of Lexical Variation. Meaning, Naming, and Context.* Berlin/New York: Mouton de Gruyter.
Geeraerts, Dirk, and Hubert Cuyckens (eds.)
    2007   *The Oxford Handbook of Cognitive Linguistics*. Oxford: Oxford University Press.
Geiger, Richard, and Brygida Rudzka-Ostyn (eds.)
    1993   *Conceptualisations and Mental Processing in Language*. Berlin/New York: Mouton de Gruyter.
Gries, Stefan Th.
    2006   Corpus-based methods and cognitive semantics: the many meanings of *to run*. In *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*, Stefan Th. Gries, and Anatol Stefanowitsch (eds.), 57-99. Berlin/New York: Mouton de Gruyter.
Gries, Stefan Th., and Dagmar Divjak
    2009   Behavioural profiles: a corpus-based approach to cognitive semantic analysis. In *New Directions in Cognitive Linguistics*. Vvyvan Evans, and Stéphanie Pourcel, 57-75. Amsterdam: Benjamins.
Gries, Stefan Th., and Anatol Stefanowitsch (eds.)
    2006   *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*. Berlin/New York: Mouton de Gruyter.
Hasan, Ruqaiya
    1989   Semantic variation and sociolinguistics. *Australian Journal of Linguistics* 9: 221–275.
    1992   Meaning in sociolinguistic theory. In *Sociolinguistics Today: International Perspectives*, Kingsley Bolton, and Helen Kwok (eds.), 80-119. London: Routledge.
    2009   *Semantic Variation: Meaning in Society. (Collected Works of Ruqaiya Hasan,* ed. by Jonathan J. Webster*)*. London: Equinox.

Hosmer, David W., and Stanley Lemeshow
    1989   *Applied Logistic Regression*. New York: John Wiley and Sons.
Kerswill, Paul
    2006   *Socio-economic class*. In *The Routledge Companion to Sociolinguistics*, Carmen Llamas, Louise Mullany, and Peter Stockwell (eds.), 51-61. London: Routledge.
Kleinbaum, David
    1994   *Logistic Regression: A Self-Learning Text*. New York: Springer.
Kristiansen, Gitte, and René Dirven (eds.)
    2008   *Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems*. Berlin/New York: Mouton de Gruyter.
Kurath, Hans (ed.)
    1941   *Linguistic Atlas of New England.* Providence, RI: Brown University.
Labov, William
    1963   The social motivations of a sound change. *Word* 19: 273- 309.
    1966   *The social stratification of English in New York City*. Washington: Center for Applied Linguistics.
    1972   *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.
    1994   *Principles of Linguistic Change*. Oxford: Blackwell.
Lakoff, George
    1987   *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago: University of Chicago Press.
Llamas, Carmen, Louise Mullany, and Peter Stockwell (eds.)
    2007   *The Routledge Companion to Sociolinguistics*. London: Routledge.
Macaulay, Ronald K. S.
    2005   *Talk that Counts: Age, Gender, and Social Class Differences in Discourse*. Oxford/New York: Oxford University Press.
    2006   Pure grammaticalization: the development of a teenage intensifier. *Language Variation and Change* 18 (3): 267-283.
Moss, Marc, Andrew D. Wellman, and George A. Cotsonis
    2003   An appraisal of multivariable logistic models in the pulmonary and critical care literature. *Chest* 123 (3): 923-928.
Nerlich, Brigitte, and David D. Clarke
    1992   Semantic change: case studies based on traditional and cognitive semantics. *Journal of Literary Semantics* 21: 204- 225.
Orton, Harold, and Eugen Dieth (eds.)
    1962- 71 *Survey of English Dialects*. Leeds: Arnold.
Peng, Chao-Ying Joanne, Kuk Lida Lee, and Gary M. Ingersoll
    2002   An introduction to logistic regression analysis and reporting. *The Journal of Educational Research* 96 (1): 3-14.

Peters, Hans
    1988   On the state and aims of Middle English word geography. In *Historical Dialectology. Regional and Social.* Trends in Linguistics. Studies and Monographs 37, Jacek Fisiak (ed.), 397- 416. Berlin/New York: Mouton de Gruyter.

Pope, Jennifer
    2002   Revisiting Martha's Vineyard. M.A. thesis, University of Edinburgh.

Poplack, Shana, David Sankoff, and Christopher Miller
    1988   The social correlates and linguistic processes of lexical borrowing and assimilation. *Linguistics* 26: 47- 104.

Roberts, Kenneth
    2001   *Class in Modern Britain*. Basingstoke, UK: Palgrave Macmillan.

Romaine, Suzanne
    1984   *The Language of Children and Adolescents: Acquisition of Communicative Competence*. Oxford, Blackwell.

Rosch, Eleanor
    1975   Cognitive representations of semantic categories. Journal of Experimental Psychology 104: 199- 233.

Ruiz de Mendoza Ibáñez, Francisco José, and M. Sandra Pena Cervel (eds.)
    2005   *Cognitive Linguistics: Internal Dynamics and Interdisciplinary Interaction.* Berlin/ New York: Mouton de Gruyter.

Sankoff, Gillian
    2006   Age: apparent time and real time. *Elsevier Encyclopedia of Language and Linguistics*, Keith Brown (ed.), 1: 110-116. Amsterdam, Elsevier.

Stenström, Anna-Brita
    2000   It's enough funny, man: intensifiers in teenage talk. In *Corpora Galore: Analyses and Techniques in Describing English*, John M. Kirk (ed.), 177-190. Amsterdam: Rodopi.

Tabachnick, Barbara G., and Linda S. Fidell
    2001   *Using Multivariate Statistics*. Boston: Allyn and Bacon.

Tagliamonte, Sali A., and Alex D'Arcy
    2004   He's like, she's like: the quotative system in Canadian youth. *Journal of Sociolinguistics* 8 (4): 493-514.
    2009   Peaks beyond phonology: adolescence, incrementation, and language change. *Language* 85 (1): 58-108.

Trudgill, Peter
    1974   *The Social Differentiation of English in Norwich*. London: Cambridge University Press.

Upton, Clive, and John Widdowson
    1999   *Lexical Erosion in English Regional Dialects*. Sheffield: The University of Sheffield.

Weinreich, Uriel, William Labov, and Marvin I. Herzog
    1968   Empirical foundations for a theory of language change. In *Directions for Historical Linguistics: A Symposium*, Winfred P. Lehmann, and Yakov Makiel (eds.), 97-195. Austin: University of Texas Press.

Wong, Andrew
    2002   The semantic derogation of Tongzhi: a synchronic perspective. *Language and Sexuality: Contesting Meaning in Theory and Practice,* Kathryn Campbell-Kibler, Robert J. Podesva, Sarah J. Roberts, and Andrew Wong (eds.), 161-174. Stanford, CA: CSLI.
    2008   On the actuation of semantic change: the case of Tongzhi. *Language Sciences* 30 (4): 423–449.

Applying word space models to sociolinguistics. Religion names before and after 9/11.

Yves Peirsman, Kris Heylen and Dirk Geeraerts

Abstract

Researchers in disciplines like lexical semantics and critical discourse analysis are in need of a quantitative method that allows them to model the distribution of a word automatically. We advocate the use of word space models, a family of approaches that were developed in the context of computational linguistics and cognitive science, which represent the meaning of a word in terms of its contexts in a large corpus. In a case study on the use of religious terms before and after the attacks of September 11, 2001, we show how these models can be employed to determine the semantic similarity and relatedness between two words, and the factors that influence them. One of the patterns we uncover is the increased association between Islam and terrorism in Dutch newspaper articles after 9/11, a trend that is far less outspoken for Christianity. We also apply these new quantitative instruments to explore the differences in word use between the five newspapers in our corpus, and find a striking distinction between popular and quality newspapers.

Keywords: lexical semantics, word space models, semantic similarity, association, religious terms, changes in word use

1. Introduction

Of all computational-linguistic approaches to lexical semantics, word space models currently set the trend (see e.g., Padó and Lapata 2007). Based on the hypothesis that semantically similar words tend to be used in similar contexts, these corpus-based approaches model the meaning of a word in terms of the contexts in which it appears. They are applied to a wide variety of computational tasks – from Question Answering and Information Retrieval to automated essay scoring (Landauer and Dumais 1997) or the modeling of human behavior in psycholinguistic experiments (Lowe and McDonald 2000). In this article, we will argue that such word space models

can be equally useful to more theoretically-oriented linguistic research. Thanks to their fully automatic analysis of the distribution of a word, word space models are not only able to deal with enormous quantities of data; they also bypass the need for subjective human judgments and may bring to light patterns that escape the human eye.

   In the next section, we will argue that the time is right for word space models to be introduced into theoretical and descriptive (socio-)linguistics. We will present two types of models that are often used in computational linguistics – the document-based and syntax-based approaches – and show how they work in practice. In section 3, we will then illustrate our case with a variational-linguistic study. In particular, we will investigate how the use of the religion names *islam* 'Islam' and *christendom* 'Christianity' has changed after the attacks of 11 September 2001. To this goal, we study a large Dutch corpus of about 300 million words, consisting of newspaper articles from between 1999 and 2002. In section 4, finally, we wrap up with conclusions and an outlook for future research.

2. Word space models of lexical semantics

2.1. Related work

Despite the fact that word space models are mainly investigated in the field of computational linguistics, their origin lies in linguistic insights. Through the history of linguistics and language philosophy, a number of researchers have stressed the dependency, or even the identity, between the meaning of a word and its use. This view inspired John R. Firth's (1957) quote that "you shall know a word by the company it keeps", Ludwig Wittgenstein's "the meaning of a word is its use in the language" (1953, p. 43), and Zellig Harris' (1954) insight that semantically similar words are used in similar contexts – a view which is now often referred to as the *distributional hypothesis*. However, in its quest for models of word meaning, theoretical linguistics has embraced these views much less enthusiastically than the more applied disciplines.

   It might be argued, however, that the time has come for such word space models to find their way to more theoretically-oriented research. Not only are computational linguists fast gaining insight in the semantic characteris-

tics of these models (Sahlgren 2006, Padó and Lapata 2007, Peirsman, Heylen and Speelman 2007, Heylen, Peirsman, Geeraerts and Speelman, 2008); in recent years, a number of related approaches in corpus linguistics have also paved the way. In line with its increased interest in corpora (see e.g., Tummers, Heylen, and Geeraerts 2005), the cognitive-linguistic community in particular has grasped the importance of a usage-based study of lexical semantics, based on more advanced techniques than just the extraction of examples from corpora. Such corpus-based approaches to lexical semantics are the focus of a number of recent anthologies (e.g. Gries and Stefanowitsch 2006, Stefanowitsch and Gries 2006) and were the topic of a successful theme session at the 10th International Cognitive Linguistics Conference. Advanced statistical methods, like clustering techniques or correspondence analysis, are currently at the centre of attention.

Basically, there are two perspectives such a corpus-based study of lexical semantics can take. First, it is possible to focus on one polysemous word, and investigate the syntactic or lexical features that correlate with the occurrences of its several meanings. This semasiological approach is represented by Gries' (2006) study of the English verb *to run*. Gries labels all occurrences of *to run* in ICE-GB and the Brown Corpus with a number of tags that together form the *behavioral profile* of the verb. This profile contains morphological features, syntactic properties of the clause, semantic characteristics of the relevant participants, collocates of the verb in the same clause and a paraphrase of the verb's meaning. Gries then uses this data to identify the distinct senses of *to run*, to find its prototypical sense, and to determine how these can be combined in a network, among others. Basically, this approach is a computational alternative to the traditional work of a lexicographer or lexicologist: it tries to identify the contexts that go together with the specific senses of a word (Geeraerts 1997).

Second, it is also possible to study not just one word, but a set of words and the differences and similarities between them. This more onomasiological perspective is taken by Divjak and Gries (2006), who cluster verbs of *trying* in Russian according to their behavior in a corpus. Similarly, Glynn (2009) explores the differences in behavior between the verbs *annoy*, *bother* and *hassle* in British and American English.

Word space models of lexical semantics allow for both types of investigation. On the one hand, they can be used to cluster the various occurrences of a word into groups that often largely correspond to the several senses of that word (Schütze 1998). On the other, they make it possible to find the similarities between several words on the basis of their contexts in a corpus.

Nevertheless, these approaches model the behaviour of a word quite differently from the analyses above: word space models look only at the *surface context* of a target word – defined in terms of articles or paragraphs, context words, or syntactic relations. As a result, they can do without any kind of manual labelling, and drastically increase the number of data we can deal with. This computational approach to lexical semantics can therefore provide a useful quantitative tool in fields like variational linguistics, or Critical Discourse Analysis (CDA).

In short, while word space models may be new to the study of lexical semantics, they have predecessors in the form of behavioural profiles, quantitative onomasiological analyses and the time-honoured method of manual lexicographic description. These advanced corpus-based techniques have created the right atmosphere for the introduction of word space models in variational-linguistic research, or in fields that generally bring forth more qualitative studies, like Critical Discourse Analysis.

2.2. Computational background

In computational linguistics, word space models of lexical semantics have been around for quite a while now. In the literature a wide variety of approaches has been developed and discussed (see Schütze 1998, Lin 1998, Purandare and Pedersen 2004, Sahlgren 2006, Padó and Lapata 2007 and many others). The earlier models are often still the most popular ones, with Latent Semantic Analysis (LSA, Landauer and Dumais 1997) and the Hyperspace Analogue to Language (HAL, Lund and Burgess 1996) as the two most well-known examples.

Despite all these different implementations, all word space models have the same goal: to approximate word meaning by modeling word use. They do this by keeping track of the contexts in which a word appears. In our case study below, we will make use of two types of word space models: a *document-based* and a *syntax-based* approach. Document-based models (Landauer and Dumais 1997) express the distribution of a word in terms of the articles (documents) in which it appears. Two words are thus related if they often appear in the same articles. A syntax-based model, by contrast (Lin 1998), defines the context of a target word as the context words with which it is syntactically related, plus the type of syntactic relation involved. Here two words are related if they often fulfill the same syntactic role or function in a sentence.

These two models tend to find totally different types of semantic relatedness. Syntax-based models have proved to be most accurate and efficient in the calculation of semantic similarity (Lin 1998, Padó and Lapata 2007, Peirsman, Heylen, and Speelman 2007). They generally model *paradigmatic* relations between words, like synonymy, hyponymy or hyperonymy. Document-based models are better geared towards the modeling of *syntagmatic* relations, as between *doctor* and *hospital* or *car* and *drive* (Sahlgren 2006).

In order to make this more concrete, let us take a look at the workings of a small syntax-based model. Suppose we have at our disposal a large, syntactically analyzed corpus of English, which contains the following sentences:

> Every day, he has a glass of red *wine* before he goes to bed.
> I drank too much *wine* yesterday.
> I brought home twenty bottles of red *wine* from France.

> Men drink more *beer* than women.
> I gave him twelve bottles of Belgian *beer* for his birthday.
> He has sworn to never drink *beer* again.

> She bought a new *car* last year.
> My mum prefers red *cars* to blue ones.
> I parked my *car* a few blocks from your flat.

Say now we are interested in the semantic relatedness between the target words *wine*, *beer* and *car*. Obviously, we would like to find that *beer* and *wine* are more semantically related to each other than to *car*. This can be done by studying the behavior of each target word with respect to a number of pre-defined *contextual features*. In theory, contextual features can be any characteristic of the context that may be relevant to the meaning of the target word. Semantically 'empty' words like *have* or *my* are therefore often ignored. For instance, we might count how often our target words appear in a specific syntactic relation. In this way we will mostly find paradigmatic relationships. We store our figures in long lists of frequencies – one for each target word. These 'lists' are referred to as *context vectors*, and the contextual features are their *dimensions*. For our three target words, and nine syntactic features, our toy corpus gives the following context vectors:

	d.obj_drink	d.obj_buy	d.obj_prefer	d.obj_park	pp.obj_glass_of	pp.obj_bottle_of	has.adj_red	has.adj_Belgian	has.adj_new
wine	1	0	0	0	1	1	2	0	0
beer	2	0	0	0	0	1	0	1	0
car	0	1	1	1	0	0	1	0	0

Figure 1. Syntax-based context vectors of the words *wine*, *beer*, and *car*

They indicate, for instance, that *wine* occurs as the direct object of *drink* once, *beer* twice and *car* never. In reality, of course, the vectors will have far more dimensions and often much higher values than 1 or 2. However, even on the basis of this simple example, it is clear that the two most similar context vectors are those of *wine* and *beer*. This is confirmed when we compute the quantitative similarity between the two vectors. In the literature on word space models, the most popular approach is to calculate the cosine of the angle described by the two vectors (see e.g., Bullinaria and Levy 2008). This metric gives us the following figures:

$$\cos(wine, beer) = 0.46$$
$$\cos(wine, car) = 0.38$$
$$\cos(beer, car) = 0.0$$

We have now reached the desired outcome: *wine* and *beer* are indeed more paradigmatically related to each other than to *car*. The word pair *wine – car* also has a non-zero cosine value, because both words appear with *red* as a modifying adjective. *Beer* never does.

Of course, we need not use syntactic relations as contextual features. Latent Semantic Analysis, for instance (Landauer and Dumais 1997), ignores syntax altogether. Instead, it counts the number of times each target word occurs in the documents that make up the corpus. For a newspaper corpus, we may want to determine the frequencies of our target words in each of the thousands of articles, for instance. These articles then form the dimensions of the context vectors, in the same way as the syntactic relations above. The semantic relatedness between two target words is now again

quantified as the cosine of the angle between their context vectors. Because *wine* and *beer* will probably occur together in more articles than *beer* and *car*, we will again find a higher semantic relatedness between the former word pair. Indeed, both on the syntagmatic and the paradigmatic axis, *wine* and *beer* are closer to each other than they are to *car*.

In short, word space models allow us to approximate our intuitions about the semantic relationship between two words by simply modeling the meaning of the words in terms of the contexts in which they occur. It goes without saying that the computational implementation goes slightly further than our sketch here. For instance, context vectors in practice rarely use the raw co-occurrence frequencies of the features and the target word, since these are heavily dependent on the nature of the features. Syntactic relations that occur very frequently in the corpus (syntax-based models) or extremely long articles (document-based) will automatically have high values for a large number of target words. This problem is usually tackled by replacing the raw frequencies with a statistical measure like point-wise mutual information, which indicates if the target word and the feature occur together more or less often than we expect on the basis of their individual frequencies. For these and other technical details, we refer the interested reader to our more computationally-oriented papers (Peirsman, Heylen and Speelman 2007, Heylen, Peirsman, Geeraerts and Speelman, 2008).

2.3. Case study

So far, word space models have proved their usefulness mainly in the field of computational linguistics. It is our belief, however, that they can equally be applied to more theoretical-linguistic research questions. Like other advanced empirical approaches, they have the major advantage that they can cope with far more examples than any manual analysis can, and that they can help identify patterns that would otherwise remain hidden from the human eye.

In this paper, we will apply word space models to an investigation of language variation. In particular, we will focus on a corpus of newspaper text, and try to find out in what way the use of religion names, particularly *islam* 'Islam' and *christendom* 'Christianity', has changed after the terrorist attacks of September 11, 2001. Through the empirical investigation of the contexts in which these words are used, we will pin down changes in typical contexts and hence, shifts in media coverage. Our investigation can thus

be situated at the crossroads of diachronic linguistics, sociolinguistics and Critical Discourse Analysis.

The background of this study is probably well-known. Ever since 11 September 2001, there is increased talk of a so-called clash of civilizations, a cultural opposition between East and West – Islam and Christianity – which lies at the basis of the September 11 attacks. This clash of civilizations was first introduced by Samuel Huntington in a 1993 article (Huntington 1993) and later developed in a number of books (see e.g., Huntington 1996). After 9/11, it is often argued, this view has governed international news coverage, particularly in the US media (see e.g., Abrahamian 2003, Seib 2004). Relevant studies, however, are generally of anecdotal nature. Abrahamian (2003), for instance, merely illustrates his argument with a number of well-chosen quotes from a variety of newspapers. Chang and Mehan (2006) analyze speeches by President Bush, including interviews, in a more systematic manner. They claim the existence of a War on Terror *script*, which is dominated by a religious mode of representation.

Such qualitative analyses of the discourse of power are typical products of Critical Discourse Analysis. "Critical discourse analysis (CDA) is a type of discourse analytical research that primarily studies the way social power abuse, dominance, and inequality are enacted, reproduced, and resisted by text and talk in the social and political context" (Van Dijk 2001: 352). Often, CDA tries to uncover power relations through which one group of people may control another. One such powerful group is that of journalists or editors of newspapers, who thank their influence to the knowledge and information they spread through their articles. Through their access to this media discourse, they may influence the way other people think about certain issues. This can happen in a number of ways – through the choice for or against news coverage of specific events, or, at a lower level, through a choice for specific words rather than others in their articles (Van Dijk 2001). So far, however, CDA analyses of political discourse or media texts are mainly of a qualitative nature (see e.g., Chilton 2004, Fairclough 1995).

Other fields of linguistics, too, analyze discourse mainly from this qualitative perspective. Martin (2004), for instance, discusses one particular editorial about the consequences of 9/11 in the framework of systemic-functional linguistics. In highlighting specific words and constructions, he shows what rhetorical devices the author uses to negotiate solidarity with his readership, among other elements.

Whereas such detailed analyses of individual texts may indeed bring to light important patterns and trends, other questions can only be answered

by a more quantitative approach. For instance, if we want to find out whether the use of words like *Islam* in newspapers has really changed after 9/11, we cannot satisfy ourselves with focusing on the handful of articles that a manual analysis is necessarily restricted to. In order to uncover the larger trends or differences in habit, some of which may even happen subconsciously, we need to make use of a more advanced linguistic apparatus. This can be done by studying the words that are used together with *Islam*. As Dunn, Moore and Nosek (2005) point out, these *context words* can dramatically change people's perception of the incident that is being described. Even subtle differences like substituting *strategy* for *plot* can contribute to people's opinion of an action, as either terrorist or patriotic. We believe that the framework of distributional semantics can help us identify such underlying changes and trends. Because word space models can neatly deal (and indeed, only work well) with large quantities of data, they allow a detailed linguistic investigation of word use in large corpora.

3. Case study

3.1. Experimental setup

As we indicated above, we will use two types of word space models for our analyses. The document-based model will mainly give us information on the *syntagmatic* relations of our target words. These often capture the general topic of the articles that the target words appear in. The syntax-based model will give us more *paradigmatically* related words. These are words that often have the same syntactic function or role as the target word. In total, we take eight types of syntactic relations into account:

- subject of verb *v* (su/v)*,*
- direct object of verb *v* (obj1/v) *,*
- prepositional complement to verb *v* introduced by preposition *p* (pc/v_p),
- head of an adverbial PP to verb *v* introduced by preposition *p* (advPP/v_p),
- modified by adjective *a* (mod/a),

- postmodified by a PP with head *n*, introduced by preposition *p* (pmPP/n_p),
- modified by an apposition with head *n* (app/n), or
- coordinated with head *n* (cnj/n).

Each specific instantiation of the variables *v*, *p*, *a*, or *n* leads to a new context feature. For each feature in both models, the context vector of a target word contains its point-wise mutual information with that target. As above, the similarity between two vectors is calculated on the basis of the cosine of the angle between them.

Our data was the Twente Nieuws Corpus, a corpus with 300 million words of Dutch newspaper text from between 1999 and 2002. The corpus was developed at the University of Twente and parsed by the Alpino parser at the University of Groningen. We divided the material into two subcorpora. The first contained all articles up to August 2001, the other started from October 2001. This allowed us to contrast word use before and after the terrorist attacks of September 11, 2001.

3.2. Global word use before and after September 11, 2001

The question that concerns us here is whether the use of the religion name *islam* has changed after the attacks of September 11, 2001. Intuitively, most people would argue it has indeed. It is often felt that Islam as a religion is now more clearly linked to terrorism than it was before and, as a result, is often covered in a more negative way by the news media. These intuitions, however, are in need of an empirical foundation. Only a thorough usage-based linguistic study can show if indeed the link between Islam and terrorism has become clearer and additionally, if this is also true for other religions like Christianity. For such a large-scale study, the analytical tools of Critical Discourse Analysis do not suffice anymore. We will therefore use our word space models to map the use of the words *islam* 'Islam' and *christendom* 'Christianity' in the pre-9/11 and post-9/11 parts of our corpus, both in terms of syntax-based and document-based distribution. We will then study the results, both in a diachronic (before and after 9/11) and a synchronic (*islam* vs *christendom*) way. Finally, we will investigate if there are any differences between the individual newspapers.

3.2.1. Syntax-based distribution

One way of determining how the distribution of *christendom* and *islam* has changed is to look at the words in the corpus with the most similar syntax-based distribution to either of those. Therefore we compared the context vectors of *christendom* and *islam* to those of the 10,000 most frequent nouns in the corpus, and selected from those the 100 nouns with the most similar context vectors. We refer to these nouns as the 100 *nearest neighbors* of *christendom* and *islam*. A comparison of these lists, both between the two subcorpora and the two target words, brings to light some interesting differences.

First we want to find out in what way the use of *islam* and *christendom* has changed after 9/11, and whether we see an increased similarity with *terrorisme* 'terrorism', for instance. We therefore contrast the lists of nearest neighbors to *islam* and *christendom* before and after 9/11. For each nearest neighbor, we calculate its difference in ranks between the two subcorpora, in order to discover which neighbors have climbed on the list, and which ones have fallen. If a nearest neighbor does not appear in one of the lists, it is automatically assigned rank 101 for that list. Moreover, instead of using the original ranks, we compute the difference between the natural logarithm of the ranks. This logarithmic scale ensures that differences far down in the list of nearest neighbors are treated as less important than those at the top: for instance, we want the difference between 1 and 20 to be much larger than that between 81 and 100.

Let us give an example. *Moslim* 'Muslim' was the 16th nearest neighbor to *islam* before 9/11, but climbs to 6th place afterwards. Its difference score is therefore ln(6)-ln(16) = 0.981. *Koran*, the 19th nearest neighbor to *islam* after 9/11 does not appear in the list before 9/11. Its difference score is therefore ln(19)-ln(101) = 1.671. Calculated thus, the ten highest climbers of *islam* and *christendom* are given in Table 1. The eyecatcher of Table 1 is *terrorisme* 'terrorism': the highest climber of *islam* (position 12 after 9/11) is only the 11th highest climber for *christendom* (position 50 after 9/11). The table also shows a tighter link between *fundamentalisme* 'fundamentalism' and both religions after 9/11, and between *jihad* 'jihad' and *islam*. The other highest climbers are either more neutral in meaning or display an expected link with either of the two religions (e.g., *Koran* 'Quran' and *islam*). In short, there is indeed a notable increase in syntax-based relatedness between *islam* and a number of words related to terrorism. This increase is far less clear with *christendom*.

Table 1. The ten highest climbers in the list of nearest neighbors to islam and christendom after 9/11 for the syntax-based model

islam	christendom
terrorisme 'terrorism'	*rechtsstaat* 'constitutional state'
koran 'Quran'	*jodendom* 'Judaism'
moderniteit 'modernity'	*burgerschap* 'citizenship'
islamist 'islamist'	*moderniteit* 'modernity'
verzorgingsstaat 'welfare state'	*hindoeïsme* 'Hinduism'
religie 'religion'	*christen* 'Christian'
fundamentalist 'fundamentalist'	*moraal* 'moral'
islamiet 'Islamite'	*protestante* 'Protestant'
moslim 'muslim'	*fundamentalisme* 'fundamentalism'
jihad 'jihad'	*idealism* 'idealism'

Obviously, there still exists considerable overlap between the neighbor lists of the two religions. It would therefore be interesting to look for the nearest neighbors that are typical of one religion or the other. We can do this again by calculating the difference in ranks, but this time we contrast the post-9/11 *islam* list with the post-9/11 *christendom* list. Table 2 shows the neighbors with the highest scores, i.e. the items that display the biggest differences between their rank in the *islam* list and their rank in the *christendom* list.

Let us investigate the results in detail. The most highly distinguishing neighbor for *islam* is *christendom*, and vice versa. This is no surprise, since a word cannot have itself as its nearest neighbor, and the two religions are obviously semantically related. Apart from the expected differences, like *moslim* 'Muslim' or *Koran* 'Quran' to the left and *christen* 'Christian' or *kerk* 'church' to the right, the two lists show a surprising internal consistency. On the one hand, most neighbors with a clear preference for *islam* are of an outspoken political nature. They either directly refer to politics (*verzorgingsstaat* 'welfare state', *regime* 'regime', *democratie* 'democracy', *politiek* 'politics') or to issues high on the political agenda in many countries (*minderheid* 'minority', *homoseksualiteit* 'homosexuality', *immigratie* 'immigration', *integratie* 'integration'). In addition we find a few words related to terrorism (*terrorisme* 'terrorism', *terreur* 'terror'). On the other hand, most neighbors typical of *christendom* are of a much more cultural, artistic or social nature. Here we see words related to faith (*spiritualiteit* 'spirituality', *theologie* 'theology', *mystiek* 'mysticism', *jodendom* 'Ju-

daism', *hindoeïsme* 'Hinduism', *ethiek* 'ethics', *protestante* 'Protestant', *boeddhisme* 'Buddhism', *individualisme* 'individualism', *idealisme* 'ideal- ism') and art (*literatuur* 'literature', *schilderkunst* 'painting'), along with some more socio-political movements (*nationaal-socialisme* 'National So- cialism', *sociaal-democratie* 'social democracy', *socialisme* 'socialism'). We may conclude that, when the use of *islam* and *christendom* is compared in newspaper articles after 9/11, *islam* scores particularly high on the politi- cal dimension, while *christendom* is more related to the socio-cultural sphere.

Table 2. Most highly distinguishing nearest neighbors of *islam* and *christendom*

islam	christendom
christendom 'Christianity'	*Islam* 'Islam'
moslim 'Muslim'	*spiritualiteit* 'spirituality'
Koran 'Quran'	*theologie* 'theology'
terrorisme 'terrorism'	*nationaal-socialisme*
minderheid 'minority'	'National Socialism'
islamist 'Islamist'	*modernisme* 'modernism'
religie 'religion'	*sociaal-democratie* 'social democracy'
fundamentalist 'fundamentalist'	*mystiek* 'mysticism'
jihad 'jihad'	*christen* 'Christian'
verzorgingsstaat 'welfare state'	*jodendom* 'Judaism'
denkbeeld 'conception'	*hindoeïsme* 'Hinduism'
fundamentalisme 'fundamentalism'	*literatuur* 'literature'
regime 'regime'	*ethiek* 'ethics'
homoseksualiteit 'homosexuality'	*schilderkunst* 'painting'
Islamiet 'Islamite'	*protestante* 'Protestant'
immigratie 'immigration'	*boeddhisme* 'Buddhism'
terreur 'terror'	*individualisme* 'individualism'
democratie 'democracy'	*kerk* 'church'
politiek 'politics'	*socialisme* 'socialism'
integratie 'integration'	*idealisme* 'idealism'
	mensheid 'human nature'

How can we explain this pattern in linguistic terms? What is it that makes *islam* so similar to *terrorisme* and *christendom* to *spiritualiteit*? In order to

answer this question, we have to turn our attention to the contextual features that we used as dimensions of the context vectors. We contrast *islam* and *christendom* in the same way as before, but we now replace the nearest neighbors by these contextual features. Each feature is assigned a rank on the basis of its value for the respective target word, again with rank 101 for all features outside the top 100. Table 3 gives the features with the biggest difference in log ranks after 9/11.

Table 3. Most highly distinguishing dimensions of *islam* and *christendom* after 9/11

islam	christendom
cnj/christendom/noun 'christianity'	*cnj/islam/noun* 'islam'
mod/politiek/adj 'political'	*su/heb/verb* 'have'
mod/radicaal/adj 'radical'	*mod/vroeg/adj* 'early'
cnj/moslim/noun 'muslim'	*cnj/humanisme/noun* 'humanism'
cnj/Westen/name 'West'	*mod/westers/adj* 'western'
su/sta/verb 'stand'	*cnj/hindoeïsme/noun* 'Hinduism'
cnj/democratie/noun 'democracy'	*cnj/kerk/noun* 'church'
cnj/integratie/noun 'integration'	*su/ga/verb* 'go'
mod/fundamentalistisch/adj 'fundamentalist'	*cnj/god/noun* 'god'
	cnj/boeddhisme/noun 'Buddhism'
obj1/misbruik/verb 'abuse'	*su/heers/verb* 'rule'
mod/militant/adj 'militant'	*mod/traditioneel/adj* 'traditional'
cnj/cultuur/noun 'culture'	*cnj/religie/noun* 'religion'
cnj/terrorisme/noun 'terrorism'	*mod/orthodox/adj* 'orthodox'
pc/weet_van/verb 'know_of'	*mod/protestants/adj* 'protestant'
su/vorm/verb 'form'	*su/maak_door/verb* 'go_through'
cnj/homoseksualiteit/noun 'homosexuality'	*pmPP/in_eeuw/noun* 'in_century'
	cnj/jodendom/noun 'Judaism'
mod/Nederlands/adj 'Dutch'	*pc/ga_over_tot/verb* 'proceed_to'
su/verbied/verb 'forbid'	*su/ontsta/verb* 'arise'
pmPP/in_Nederland/name 'in_Holland'	
mod/liberaal/adj 'liberal'	

This top twenty of features most typical of either *islam* or *christendom* again shows some intriguing patterns. With *christendom*, it is the religious

dimensions that dominate. The word is often conjoined with other nouns like *islam* 'Islam', *humanisme* 'Humanism', *jodendom* 'Judaism', *hindoeïsme* 'Hinduism', etc. Moreover, all modifying adjectives highly typical of *christendom* refer to the history of Christianity (*vroeg* 'early', *traditioneel* 'traditional') or to its several subgroups (*orthodox* 'orthodox', *protestants* 'protestant'). Islam, by contrast, is characterized by features that refer to politics. Again these are mainly conjunctions – with *democratie* 'democracy' and *integratie* 'integration', for instance – and modifying adjectives – like *politiek* 'political' or *radicaal* 'radical'. These lists of twenty most distinguishing dimensions thus explain why the nearest neighbors typical of *islam* tended to be of a political nature, while those of *christendom* were more oriented towards religion and culture.

3.2.2. Document-based distribution

The syntax-based model has singled out those words in the corpus with the syntactic profile most similar to that of *islam* or *christendom*. We will now ask a second question: do *islam* and *christendom* crop up in different types of articles after the September 11 attacks? For instance, we might expect *islam* to occur more often in articles on terrorism than it used to do. This change might be absent from *christendom*, or it might apply to religion names in general. To answer this second type of question, we have modeled the use of *islam* and *christendom* with a document-based word space model. As a start, we again calculated the difference in nearest neighbors, first between the pre-9/11 and the post-9/11 corpus and then between *islam* and *christendom*, as above.

Although the words are different, Table 4 mirrors the patterns we also observed above. Much more often than before 9/11, *islam* is now related with negative words like *achterlijk* 'backward', *radicaal* 'radical' or *bedreiging* 'threat'. The presence of *11* in this list moreover suggests a direct relationship to the September 11 attacks. Note that our word space models do not yet take multi-word units into account. Therefore *11* shows up as an individual entity in this list, and not yet as a part of the multi-word unit *11 september*. The negatively sounding nearest neighbors of *islam* are absent from the list of highest climbers to *christendom.*

This pattern also shows up when we explicitly contrast the list of nearest neighbors to *islam* and *christendom* after 9/11, like we did above for the syntax-based model. We will not give all results here, but just highlight the

most interesting patterns. The neighbors most typical of *islam* either refer to islam-related issues directly (e.g., *imam* 'imam', *koran*, 'Quran'), or have a negative connotation (e.g., *achterlijk* 'backward', *fundamentalist* 'fundamentalist (noun)'). The neighbors most typical of *christendom*, by contrast, are all of a strictly religious nature (e.g., *jodendom* 'Judaism', *geschrift* 'Scripture'). Again, this shows how Islam is now covered by the printing press as a political issue, much more so than Christianity.

Table 4. The ten highest climbers in the list of nearest neighbors to *islam* and *christendom* after 9/11 for the document-based model

islam	christendom
achterlijk 'backward'	*keizer* 'emperor'
niet-moslim 'non-muslim (noun)'	*jezuïet* 'Jesuit'
ongelovig 'unreligious'	*vervlechten* 'tie up'
westers 'Western'	*gedoopt* 'baptized'
afvallige 'backslider'	*islam* 'Islam'
bedreiging 'threat'	*pluralisme* 'pluralism'
moslimwereld 'muslim world'	*westers* 'Western (adj)'
radicaal 'radical (adj)'	*godsdienstig* 'religious'
11 '11'	*inscriptie* 'incription'
dialoog 'dialogue'	*bijbels* 'biblical

Since we do not have any meta-information about the newspaper articles, the contextual features of the document-based model are far less informative than those of the syntax-based model. They can only tell us that two words are similar because they appeared in so many articles together. However, the document-based models provide us with an interesting tool for looking into the topical relationship between two issues, like Islam and terrorism or Islam and culture. As a result, we can measure the relatedness between different text topics.

In order to find out whether Islam has become more topically related to terrorism, we do not simply want to measure the distributional similarity between the two words *islam* and *terrorisme*. A more robust approach is to define the two relevant *lexical fields*, and to measure the relation between them. First, we will define the lexical fields of Islam and Christianity on the one hand, and those of four topics whose relationship with Islam and Christianity we want to investigate: terrorism, war, religion and culture. Next, we

will measure the distance between those lexical fields as a function of the distance between the individual words, both before and after 9/11. In this way, we can study if the perception of *islam* and *christendom* has shifted with relation to these four topics.

Lexical fields are notably hard to delimit. Moreover, a manual construction may suffer from randomness or subjective judgments. We can, however, use document-based word space models to define our lexical fields automatically. A lexical field is then operationalized as the words with the tightest document-based relation to a central word like *terrorism* or *culture*. This is the approach we will take here.

There are a number of ways in which this undertaking can be carried out. First, it can be argued that lexical fields are far from stable entities. It would be no surprise if the lexical field of terrorism underwent some substantial changes after the September 11 attacks. For each of the central words *terrorisme* 'terrorism', *oorlog* 'war', *religie* 'religion' and *cultuur* 'culture' we therefore defined two fields – one on the basis of the pre-9/11 corpus and one on the basis of the post-9/11 corpus. Each time we included the 20 most related words to the central word together with that central word, without manual correction. Because we also wanted to include parts of speech different from nouns, we extended the set of possible nearest neighbors from the 10,000 most frequent nouns in the corpus to all words with a frequency of 200 or more. The top ten of words most related to *terrorisme*, for example, now looks like this:

> Before 9/11: *terrorisme* 'terrorism', *terrorist* 'terrorist', *aanslag* 'attack', *Libisch* 'Libyan', *catastrofaal* 'catastrophic', *Tsetjeens* 'Chechen', *terroristisch* 'terrorist (adj)', *kaping* 'hijack (noun)', *moslimrebel* 'Muslim rebel', *moslimextremist* 'Muslim extremist'

> After 9/11: *terrorisme* 'terrorism', *strijd* 'battle', *oorlog* 'war', *terrorist* 'terrorist', *militair* 'military', *bondgenoot* 'ally', *11* '11', *Amerikaans* 'American (adj)', *terroristisch* 'terrorist (adj)', *internationaal* 'international'

Apart from a few spurious words, these automatically collected sets of words appear very reasonable indeed. Before 9/11, the lexical field of terrorism is a mixed bag of 21 words referring to a number of political and religious issues: the relationship between Russia and Chechnia, the Taliban, Libya and Islam. After 9/11 these have disappeared and been replaced by

words more related to 9/11 itself: *11* '11', *september* 'September', *Amerikaan* 'American (noun)' and *Amerikaans* 'American (adj)'. September 11 has thus had a clear impact on the words that occur together with *terrorism* in newspaper articles.

The other sets of words turn out to be equally plausible. The lexical field of war contains words like *militair* 'military', *soldaat* 'soldier' and *leger* 'army'. After 9/11, there is moreover considerable overlap with the field of terrorism, with words like *strijd* 'battle', *terrorist* 'terrorist' and *internationaal* 'international'. The field of religion, then, brings together words like *religious* 'religious', *geloof* 'faith' and *God* 'God', while that of culture is made up of words like *cultureel* 'cultural', *kunst* 'art' and *traditie* 'tradition'. Careful inspection thus shows that these sets of words indeed form a reasonable basis for the investigation of the document-based relations of *islam* and *christendom*.

Second, it would also be interesting to study the relationship of *islam* and *christendom* to stable lexical fields, which do not evolve through time. In this way, we filter out the direct influence of newly related words like *Amerikaans* or *11*. For each of our four central words, we therefore constructed a stable lexical field, consisting of the words in the intersection between the pre-9/11 and post-9/11 lexical fields. We moreover corrected these fields manually to filter out any words directly related to either Islam or Christianity. For *terrorisme*, this resulted in a lexical field consisting of *terrorisme* 'terrorism', *terrorist* 'terrorist', *terroristisch* 'terrorist' and *aanslag* 'attack'. For *godsdienst*, we now have *godsdienst* 'religion', *religie* 'religion', *religieus* 'religious', *godsdienstig* 'religious', *geloof* 'faith', *God* 'God' and *gelovig* 'religious'. For *oorlog*, we found *oorlog* 'war', *militair* 'soldier', *soldaat* 'soldier' and *conflict* 'conflict'. The lexical field of *cultuur*, finally, is made up of *cultuur* 'culture', *cultureel* 'cultural', *multicultureel* 'multicultural', *samenleving* 'society', *kunst* 'art', *taal* 'language', *cultuurbeleid* 'culture policy', *integratie* 'integration', *traditie* 'tradition', *westers* 'western', *geschiedenis* 'history' and *wereld* 'world'.

Next, we defined the lexical fields of *islam* and *christendom* as the sets of the words *islam* 'Islam', *islamitisch* 'Islamic', *moslim* 'Muslim' and *christendom* 'Christianity', *christelijk* 'Christian (adj)', *christen* 'Christian (noun)', respectively. We chose these sets because they both consist of the noun and adjective referring to the religion itself plus the noun referring to its followers. Their results should thus be perfectly comparable.

On the basis of these lexical fields, we now want to investigate whether Islam and Christianity have become more or less closely related to the top-

ics of terrorism, war, culture and religion. The relatedness between two lexical fields was measured by calculating the average relatedness between each word in field 1 and each word in field 2. In order to calculate this relatedness between a word pair, we again worked with the logarithm of either word's rank in the list of the nearest neighbors to the other word. First we looked up the rank of word 1 in the list of nearest neighbors to word 2 and vice versa. We then took the average of the natural logarithm of these ranks to get a symmetric relatedness score. Each word pair contributed equally to the total score. Finally, by dividing the relatedness before 9/11 by that after 9/11, we can see if the words have become more closely or less closely related to the lexical field. The results are given in Table 5. Note that lower scores indicate a higher relatedness between two fields.

Table 5. Average distance to the lexical fields of terrorism, war, religion and culture

field 1	type	field 2	< 9/11	9/11 >	difference
Terrorism	evolving	Islam	6.58	4.55	1.45
		Christianity	8.43	7.03	1.20
	stable	Islam	5.70	3.69	1.55
		Christianity	8.65	6.82	1.27
War	evolving	Islam	5.96	4.65	1.28
		Christianity	7.09	6.92	1.02
	stable	Islam	5.79	4.94	1.17
		Christianity	7.00	6.97	1.00
Religion	evolving	Islam	3.26	2.93	1.11
		Christianity	3.41	3.61	0.95
	stable	Islam	3.37	2.87	1.05
		Christianity	2.63	2.50	1.04
Culture	evolving	Islam	5.87	4.25	1.38
		Christianity	6.69	5.26	1.27
	stable	Islam	5.84	5.24	1.11
		Christianity	6.41	5.92	1.08

Let us start with the field of terrorism. It is clear that the words referring to Islam are much more syntagmatically related to the lexical field of terrorism than those referring to Christianity. Moreover, this difference has be-

come more pronounced after the September 11 attacks. We see an increase in relatedness score of 45% for the evolving fields and 55% for the stable fields. Both times this increase is far larger than that for *christendom*, with 20% and 27%, respectively. Note that throughout Table 5, the increases in relatedness score are higher for the evolving lexical fields than for the stable ones. This is probably because the former often also included words clearly related to Islam or Christianity, and their proportion was generally larger after 9/11. By filtering these words from the fields, as we did for the stable sets of words, the relatedness scores are less subject to change.

As could be expected from the previous result, Islam has also moved closer to the lexical field of war. Its relatedness value after 9/11 lies 28% higher when we allow the field of war to change, and 17% higher when we keep it constant. For *christendom* this evolution is totally absent: the distance between the two lexical fields has all but remained the same.

With respect to the lexical field of *religion*, the observed changes are much smaller. Both *islam* and *christendom* move slightly closer to the stable lexical field, but they move in different directions when compared with the evolving field. Apparently, words referring to Islam in particular appear more often in articles about religion after 9/11, but this evolution is rather modest.

A much more drastic change takes place in the relationship between the words referring to our religions and the lexical field of culture. Both religions have become much more related to culture when we take the evolving fields into account (38% and 27%, respectively), but have moved only slightly when we look at the stable lexical field (11% and 8%). This discrepancy is caused by the fact that after 9/11, the evolving lexical field of culture includes words referring to Islam and Christianity directly (*christen* 'Christian (adj)', *moslim* 'Muslim (noun)', etc.). These were filtered out in the stable lexical field. Either way, it seems safe to conclude that both Islam and Christianity are now more topically related to issues relating to culture.

With one exception, all evolutions in Table 5 are of a positive nature: they show how Islam and Christendom occur more often in articles together with words referring to terrorism and war as well as culture and religion. This should come as no surprise: as a result of 9/11, the relationship between religion and terrorism, war, culture and related topics has become a much hotter topic in the media. In general, it is Islam that has undergone the major changes. While religion in general has figured more prominently in newspaper articles, Islam in particular has attracted attention. Its much closer relationship to the lexical field of terrorism is the most conspicuous

change, but it goes hand in hand with tighter links to war, religion and culture as well.

3.3. Comparing newspapers

So far we have talked about our subcorpora as two undifferentiated wholes. However, both are made up of articles from five newspapers, taken together to give the aggregated results in the sections above. Obviously, these newspapers differ in a number of respects. NRC Handelsblad is the main quality newspaper in The Netherlands, and like De Volkskrant, it targets a highly educated audience. Trouw, another quality newspaper, has a specific focus on religious and philosophical issues. Algemeen Dagblad, with a far broader target audience, is the most popular newspaper in our corpus. Het Parool, finally, is often considered a left-wing newspaper and mainly focuses on Amsterdam in its news coverage. Given these differences, it is only to be expected that the several newspapers also distinguish themselves in their use of the words under investigation – something that the combined results do not give us access to. In this final section we will therefore focus on the individual newspapers and investigate if they indeed display significant differences in their use of the terms we are investigating. As we are now working on much smaller corpora, the syntax-based model starts to suffer from data sparseness. Therefore we will continue working with the document-based model only.

As above, we calculated the relatedness between the lexical fields of terrorism, war, religion, culture and our small set of words referring directly to Islam or Christianity. This time, however, we based ourselves on the subcorpora of the five different newspapers, in order to pin down possible differences in their use of these religion names. We will focus here on the results for *islam* in the post-9/11 corpus, but obviously the same technique can be applied to *christendom* and the pre-9/11 corpus. The results are shown in Table 6.

The document-based relationship between *islam* and the field of terrorism is clearest in the newspaper Algemeen Dagblad. This is in line with its status as the most popular newspaper in our corpus. Both for the evolving and stable lexical fields, it returns the smallest distance between the two. The pattern is clearest when we look at the stable field. Here Algemeen Dagblad shows a relatedness of 3.40 between Islam and terrorism, with the scores of the other papers somewhere between 3.66 (Volkskrant) and 3.73

(NRC Handelsblad). Given that many of the words in the post-9/11 (evolving) lexical field of terrorism refer to 9/11 directly, the higher difference for the stable fields probably means that Algemeen Dagblad distinguishes itself most by often using together words referring to Islam and words referring to terrorism *in general* (as opposed to more specific events). Algemeen Dagblad also shows the highest relatedness between the lexical field of Islam and that of war. This is no surprise: not only are war and terrorism clearly related themselves; the post-9/11 (evolving) fields also show a considerable overlap in words.

Table 6. Average distance between the lexical field of islam and those of terrorism, war, religion and culture for all five newspapers

	evolving fields	stable fields
Terrorism		
Algemeen Dagblad	**4.42 (1)**	**3.40 (1)**
NRC Handelsblad	4.60 (4)	3.73 (5)
Het Parool	4.56 (3)	3.57 (4)
Trouw	4.73 (5)	3.68 (3)
Volkskrant	4.53 (2)	3.66 (2)
War		
Algemeen Dagblad	**4.49 (1)**	**4.74 (1)**
NRC Handelsblad	4.67 (4)	4.91 (3)
Het Parool	4.60 (2)	4.91 (3)
Trouw	4.75 (5)	4.89 (2)
Volkskrant	4.64 (3)	4.94 (5)
Religion		
Algemeen Dagblad	3.10 (4)	3.41 (5)
NRC Handelsblad	2.76 (2)	**2.68 (1)**
Het Parool	3.18 (5)	3.07 (4)
Trouw	**2.68 (1)**	3.02 (3)
Volkskrant	2.93 (3)	2.87 (2)
Culture		
Algemeen Dagblad	4.37 (3)	5.20 (3)
NRC Handelsblad	4.18 (2)	5.00 (2)
Het Parool	4.62 (5)	5.54 (5)
Trouw	**3.95 (1)**	**4.71 (1)**
Volkskrant	4.25 (4)	5.24 (4)

A totally different picture emerges for the lexical field of religion. Here suddenly Algemeen Dagblad and Het Parool give the lowest scores between the two topics. Instead Trouw and NRC Handelsblad display the highest document-based relatedness. As we mentioned before, these are two quality newspapers in the Netherlands. In addition, Trouw also has a specific focus on religious issues. In NRC Handelsblad words referring to Islam tend to crop up more often in articles that also contain words referring to religion in general – the stable lexical field. Trouw, by contrast, scores better on words in the evolving lexical field of religion that also refer to the more political sphere, like *meningsuiting* 'freedom of speech' or *grondrecht* 'basic right'. Finally, Trouw also shows the highest relatedness between Islam and the lexical field of culture. This time the results for the stable and evolving lexical fields perfectly mirror each other. Trouw emerges as the clear winner twice, with relatedness scores of 3.95 and 4.71.

Taken together, the results in Table 6 point at two poles in the document-based relationships of *islam*, *islamitisch* and *moslim* after the September 11 attacks. On the one hand, there are the lexical fields of terrorism and war. When we compare the five newspapers, it is Algemeen Dagblad in particular that links up these fields with Islam, by referring to them in the same article. On the other hand we have the fields of religion and culture. These fields are connected most with Islam in NRC Handelsblad and Trouw.

What do these results tell us about the news coverage of the investigated newspapers? Obviously we should not conclude that Algemeen Dagblad tends to portray Islam as a source of terrorism, while NRC Handelsblad and Trouw see it more as a cultural religion. After all, texts that deny the link between Islam and terrorism would also contribute to a high relatedness score between the two lexical fields, simply because both topics are present. The results do tell us something, however, about the different perspectives that the newspapers take. It appears that Algemeen Dagblad focuses mostly on the coverage of the news events themselves, while NRC Handelsblad and Trouw like to go into the cultural and religious background of terrorism and Islam. This indeed ties up with the supposed interests of the newspapers' target audiences.

4. Conclusions

The main goal of this article was to illustrate the use of word space models in theoretically-oriented linguistics. These word space models, which capture the semantic relationship between two words in terms of their distribution in a corpus, have now long proved their use in the field of computational linguistics. We argued, however, that they can equally be applied in the more theoretical disciplines, and function as a tool for quantitative investigations in fields like variational linguistics or Critical Discourse Analysis.

Word space models come in many flavors. Those that describe a word in terms of its syntactic relations are best suited to find paradigmatic similarity, as between synonyms. This is because they uncover words in the corpus that tend to have the same functions or roles as the target word. Those models that describe a word in terms of the articles in which it appears, are best geared towards the modeling of syntagmatic, topical relations. In this way, they can be used to automatically define lexical fields and to measure the distance of between two such fields.

By way of a case study, we have shown how these word space models can be put to practice in the field of usage-based theoretical (socio-) linguistics. In particular, we studied the use of the religion names *islam* and *christendom* in Dutch newspaper articles before and after the attacks of September 11, 2001. We were able to pin down some important changes. Both the document-based and the syntax-based model showed that *islam* has become distributionally more similar to words related to terrorism and politics. *Christendom*, by contrast, was still mainly characterized by cultural and more positive dimensions. Many of these results would be difficult to find with a manual analysis. A sociolinguistic study of word fields in the several newspapers showed significant differences in their treatment of *islam*. For Algemeen Dagblad in particular we found a high relatedness value between Islam and terrorism – due to the fact that these topics often occur together in its articles. Of the five newspapers, NRC Handelsblad and Trouw appear to focus most on the cultural and religious background of the news events.

Evidently, this article has introduced the word space models in the most concise way. For those readers interested in the technical details of the models, there is a wealth of literature in computational linguistics and cognitive science. Moreover, our case study has only given a brief sketch of the many possibilities that the models offer. Their automatic approach

opens up a dazzling range of possibilities in the study of language variation – from the comparison of several newspapers to that of different genres or language varieties. We hope that this short introduction will inspire some exciting new research in these fields.

References

Abrahamian, Ervand
    2003   The US media, Huntington and September 11. *Third World Quarterly* 24 (3): 529-544.
Bullinaria, John A., and Joseph P. Levy
    2008   Extracting semantic representations from word co-occurrence statistics: A computational study. *Behaviour Research Methods* 39: 510-526.
Chang, Gordon C., and Hugh B. Mehan
    2006   Discourse in a religious mode: the Bush administration's discourse in the War on Terrorism and its challenges. *Pragmatics* 16 (1): 1-23.
Chilton, Paul
    2004   *Analysing Political Discourse. Theory and Practice*. London: Routledge.
Divjak, Dagmar, and Stefan Th. Gries
    2006   Ways of trying in Russian: clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory* 2 (1): 23-60.
Dunn, Elizabeth W., Moriah Moore, and Brian A. Nosek
    2005   The war of the words: how linguistic differences in reporting Shape perceptions of terrorism. *Analyses of Social Issues and Public Policy* 5 (1): 67-86.
Fairclough, Norman
    1995   *Media Discourse*. London: Edward Arnold.
Firth, John R.
    1957   A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis,* John R. Firth (eds), 1-32. Oxford: Blackwell
Geeraerts, Dirk
    1997   *Diachronic Prototype Semantics*. Oxford: Clarendon Press.
Glynn, Dylan
    2009   Polysemy, syntax, and variation. A usage-based method for Cognitive Semantics. In *New Directions in Cognitive Linguistics*, Vyvyan Evans, and Stephanie Pourcel (eds.), 77-104. Amsterdam: Benjamins.

Gries, Stefan Th.
  2006   Corpus-based methods and cognitive semantics: the many meanings
               of *to run*. In *Corpora in Cognitive Linguistics: Corpus-Based Ap-
               proaches to Syntax and Lexis*, Stefan Th. Gries, and Anatol Stefa-
               nowitsch (eds.), 57-99. Berlin/New York: Mouton de Gruyter.
Gries, Stefan Th., and Anatol Stefanowitsch (eds.)
  2006   *Corpora in Cognitive Linguistics.* Berlin/New York: Mouton de
               Gruyter.
Harris, Zellig
  1954   Distributional structure. *Word* 10 (23): 146-162.
Heylen, Kris, Yves Peirsman, Dirk Geeraerts, and Dirk Speelman.
  2008   Modelling word similarity. An evaluation of automatic synonymy
               extraction algorithms. In *Proceedings of the Language Resources
               and Evaluation Conference (LREC-2008)*. Marrakech, Morocco.
Huntington, Samuel
  1993   The clash of civilizations? *Foreign Affairs* 72 (3): 22-49.
  1996   *The Clash of Civilizations and the Remaking of World Order*. New
               York: Simon and Shuster.
Landauer, Thomas, and Susan T. Dumais
  1997   A solution to Plato's Problem: the Latent Semantic Analysis theory
               of acquisition, induction and representation of knowledge. *Psycho-
               logical review* 104: 211-240.
Lin, Dekang
  1998   Automatic retrieval and clustering of similar words. In *Proceedings
               of the 36th Annual Meeting of the Association for Computational
               Linguistics and the 17th International Conference on Computational
               Linguistics (COLING-ACL-1998),* 768-774. Montreal, Canada.
Lowe, Will, and Scott McDonald
  2000   The direct route: mediated priming in semantic space. In *Proceed-
               ings of the Annual Conference of the Cognitive Science Society
               (CogSci-2000)*, 675-680. Philadelphia, PA.
Lund, Kevin, and Curt Burgess
  1996   Producing high-dimensional semantic spaces from lexical co-
               occurrence. *Behaviour Research Methods, Instruments and Comput-
               ers* 28: 203-208.
Martin, James R.
  2004   Mourning: how we get aligned. *Discourse & Society* 15 (2-3): 321-
               344.
Padó, Sebastian, and Mirella Lapata
  2007   Dependency-based construction of semantic space models. *Compu-
               tational Linguistics* 33 (2): 163-199.

Peirsman, Yves, Kris Heylen, and Dirk Speelman
    2007   Finding semantically related words in Dutch. Co-occurrences vs. syntactic contexts. In *Proceedings of the Workshop on Contextual Information in Semantic Space Models (CoSMO-2007)*, Marco Baroni, Alessandro Lenci, and Magnus Sahlgren (eds.), 9-16. Roskilde, Denmark.

Purandare, Anagha and Ted Pedersen
    2004   Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL-2004)*, Hwee Tou Ng, and Ellen Riloff (eds.), 41-48. Boston, MA.

Sahlgren, Magnus
    2006   The Word-Space Model. Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. Ph.D. diss., University of Stockholm.

Schütze, Hinrich
    1998   Automatic word sense discrimination. *Computational Linguistics* 24 (1): 97-124.

Seib, Philip
    2004   The news media and the "Clash of Civilizations". *Parameters* 34: 71-83

Stefanowitsch, Anatol, and Stefan Th. Gries
    2003   Collostructions: investigating the interactions of words and constructions. *International Journal of Corpus Linguistics* 8 (2): 209-243.

Stefanowitsch, Anatol, and Stefan Th. Gries (eds.)
    2006   *Corpus-based Approaches to Metaphor and Metonymy*. Berlin/New York: Mouton de Gruyter.

Tummers, José, Kris Heylen, and Dirk Geeraerts
    2005   Usage-based approaches in Cognitive Linguistics: a technical state of the art. *Corpus Linguistics and Linguistic Theory* 1: 225-261.

Van Dijk, Teun
    2001   Critical Discourse Analysis. In *The Handbook of Discourse Analysis*, Deborah Schiffrin, Deborah Tannen, and Heidi E. Hamilton (eds.), 352-371. Malden, MA/Oxford, UK: Blackwell.

Wittgenstein, Ludwig
    1953   *Philosophical Investigations*. Oxford: Blackwell.

Part two.
Constructional variation

The English genitive alternation in a cognitive sociolinguistics perspective

Benedikt Szmrecsanyi

Abstract

As a corpus-based inquiry into the probabilistic nature of lectal variation, the present study seeks to explore how language-external determinants of linguistic variation – real time, geography, text type – interact with language-internal determinants of linguistic variation, and in so doing shape cognitive and probabilistic grammars. The concrete empirical attention of this study will be directed toward the English genitive alternation as an instructive case study. The evidence suggests that the probabilistic grammar underlying the system of genitive choice is fundamentally the same across sampling times, geographic varieties of English, and text types. This overall qualitative stability notwithstanding, the importance of individual conditioning factors varies across different data sources, and this variability is shown to be mediated by language-external factors.

1. Introduction

As is well known, English has two grammatically overt means of expressing genitive relations, the *of*-genitive (also known as the 'Norman genitive', 'periphrastic genitive', or 'of-construction'), as in (1), and the *s*-genitive (also known as the 'Saxon genitive'), as in (2):

(1)   ... this session is helpful for all of us in that it forces us to rethink, to problematize, and to interroga**te the history of American anthropology** ... (Corpus of Spoken American English, text 1034)

(2)   While **anthropology's history** is indeed implicated in the scientific con-
      struction as race as a biological fact ... (Corpus of Spoken American Eng-
      lish, text 1034)

In modern English, the two genitives are fairly interchangeable and near-
equivalent ways of saying the same thing in a considerable number of con-
texts (Jucker 1993: 121). For example, *anthropology's history* and *the his-
tory of anthropology* are certainly close paraphrases, and it is such choice
contexts that will be in the focus of attention in the present investigation.

   Where an *s*-genitive can be paraphrased by an *of*-genitive (or vice ver-
sa), which factors bear on language users' choice? Extant research has
identified a multitude of parameters affecting the English genitive alterna-
tion. The literature suggests four major language-internal factor groups:

(i)   *Semantic and pragmatic factors*. Animate possessors attract the *s*-
      genitive, inanimate possessor attract the *of*-genitive (for instance, Al-
      tenberg 1982: 117-148); increased thematicity (i.e. text frequency) of
      the possessor NP makes usage of the s-genitive significantly more like-
      ly (Hinrichs and Szmrecsanyi 2007).

(ii) *Phonology*. A final sibilant in the possessor NP (for instance, in a plur-
      al morpheme) attracts the *of*-genitive (for instance, Hinrichs and
      Szmrecsanyi 2007).

(iii) *Processing and parsing-related factors*. Thanks to the principle of end-
      weight (Behaghel 1909/1910), longer possessor NPs prefer the *of*-
      genitive (because the *of*-genitive places the possessor second) while
      heavier possessums prefer the *s*-genitive (for instance, Quirk et al.
      1985: 1282; Biber et al. 1999: 304). It is also known that language us-
      ers tend to recycle material that they have used or heard before, a phe-
      nomenon which is often psycholinguistically motivated (cf. Szmrec-
      sanyi 2005a, 2005b, 2006). Thus, precedence of either genitive
      construction in discourse (be it written or spoken) increases the odds
      that the same genitive type will be used next time there is a choice.

(iv) *Economy-related factors*. By virtue of being "more compact" (Biber et
      al. 1999: 300), the *s*-genitive is more frequent in contexts and registers
      where the "tendency to brevity" (Dahl 1971: 172) is pivotal. There is
      also evidence that journalists favor the *s*-genitive in contexts characte-

rized by comparatively high informational/lexical density (Hinrichs and Szmrecsanyi 2007).

In addition, the genitive alternation is also sensitive to a number of language-external factors:

(v) *External factors*. In historical terms, the *of*-genitive has been the long-term incoming form, yet the *s*-genitive has bounced back during the Modern English period and is claimed to be spreading right now, especially in press language (for instance, Raab-Fischer 1995; Mair 2006). As for genre/text type stratification, more informal settings usually favor the *s*-genitive (for instance, Altenberg 1982: 284) – so, the *s*-genitive should be particularly frequent in spoken data (Rosenbach 2002: 39). In terms of geographic differences, the *s*-genitive is known to be more frequent in American English than in British English (cf., for example, Rosenbach 2003: 395-396).

The univariate impact of each of the factors mentioned above is amply documented in the literature. The aim of the present research, by contrast, is to fit a multivariate logistic regression model describing the probabilistic grammar of genitive choice, with special attention being paid to how the external factors in (v) shape and determine the factor weights of the internal factors in (i) – (iv). To address this particular issue, the present study will rely heavily on visualization techniques such as cluster analysis and multidimensional scaling. An interesting issue along these lines is the cultural-cognitive motivation driving the on-going spread of the *s*-genitive, especially in press language: is this a text-type-interdependent process such that we are witnessing a 'colloquialization of the norms of written English' (Leech and Smith 2006; Hundt and Mair 1999)? Alternatively, are we seeing a geography-related process of 'Americanization' (such that the *s*-genitive would become more frequent in British English because it is frequent in American English')? Or are we rather dealing with a process of "economization" (Hinrichs and Szmrecsanyi 2007: 469), such that journalists have to increasingly convey ever more information in ever less paper space, a constraint that would favor the more compact *s*-genitive?

In addressing these issues, and thus sketching a more complete, more realistic, and thoroughly usage-based picture of linguistic variation in space, time, and across text types, the overarching objective of this study is to explore the gradient interaction between language-internal and language-

external factors as a *cognitive* and *cultural* phenomenon that comes within the remit of cognitive sociolinguistics.

2.   Method and Data

The present study will re-examine the database and the coded dataset drawn upon in Szmrecsanyi and Hinrichs (2008). This section will loosely paraphrase their methods section.

2.1. Data

The database taps the following corpora sampling naturalistic language data:

–   The *Corpus of Spoken American English* (CSAE). The release that will be used here is composed of the installments 1 and 2 (Du Bois et al. 2000; Du Bois et al. 2003), spanning in all 41 conversations, each approximately 20–30 minutes in length. Designed primarily for conversation analytic purposes and thus sampling very conversational, unscripted and hence very informal American English, this corpus is a comparatively small one (roughly 166,000 words of running text), though it is large enough for some of the purposes of the present study.

–   The *Freiburg Corpus of English Dialects* (FRED). This corpus (see Hernández 2006; Szmrecsanyi and Hernández 2007) contains samples of dialectal speech (mainly transcribed so-called 'oral history' material) from a variety of sources. The bulk of these samples was recorded between 1970 and 1990; in most cases, a fieldworker interviews an informant about life, work etc. in former days. The informants are typically elderly people with a working-class background. Speech styles are relatively formal due to the interview situation. The subsample of FRED to be analyzed here spans ca. 1.3 million words; dialect areas included in the sample are the Hebrides, the Midlands, the North of England, Wales, the Southwest, and the Southeast (the exact composition is not of interest here, as this is not a study in dialectology).

–   The A and B sections in the *Brown family of corpora* (Brown, LOB, Frown, and F-LOB). These four corpora contain written, edited, and published Standard English. The two older corpora, Brown and LOB, represent, respectively, American and British English from the 1960s, whereas Frown and F-LOB are their 1990s updates. Thus, the quartet covers two varieties and a time span of 30 years. The corpora are all structured according to a set framework of fifteen different genre categories. In total, each corpus contains 500 text samples. At a sample size of about 2,000 words each, the four Brown corpora contain a structured dataset of four million words of running text. The present study will focus on journalistic language and therefore explore the categories 'Reportage' (A) and 'Editorial' (B) from each of the corpora, amounting to 71 samples, or roughly 142,000 words, per corpus, adding up to a total of ~568,000 words, relying on the recently completed part-of-speech-tagged versions of the corpora (see Leech and Smith 2005; Hinrichs, Waibel, and Smith 2007).

In sum, the database to be explored here comprises material from different sampling times (1960s [LOB, Brown] vs. 1990s [F-LOB, Frown] press English), different geographic varieties (American English [CSAE, Brown, Frown] vs. British English [FRED, LOB, F-LOB]), and different text types (spoken [FRED, CSAE], written press reportage [Brown-A, LOB-A, Frown-A, F-LOB-A], written press editorials [Brown-B, LOB-B, Frown-B, F-LOB-B]).

2.2. Method

All occurrences of interchangeable *s-* and *of*-genitives were manually identified in the database, i.e. each instance of an *s-* or *of*-genitive was classified according to whether the alternative construction could have been used in its place. This procedure yielded a dataset of $N = 10,450$ interchangeable genitives (CSAE: $N = 332$; FRED: $N = 1,818$; Brown: $N = 2,204$; LOB: $N = 2,019$; Frown: $N = 2,132$; F-LOB: $N = 1,945$).

While Szmrecsanyi and Hinrichs (2008) provide a detailed description of the coding scheme for interchangeability, suffice it to say here that the coding procedure only considered those instances of the *s*-genitive which could equally have been expressed as an *of*-genitive by applying a simple conversion rule, without adding or deleting any of the lexemes in the possessor or possessum phrase (except for the optional addition of a determiner

to the possessum). Similarly, only those *of*-genitive tokens were retained which could have been expressed using an *s*-genitive construction instead with neither of the noun phrases modified, except for the necessary deletion of any determiner in the possessum phrase. Crucially, the alternative construction would have to leave the meaning of the actual choice unchanged; consequently, the *city of Atlanta* was not considered an interchangeable genitive because the alternative, *Atlanta's city*, has a different meaning. A negative list of non-interchangeable genitive types – roughly following the similar lists in Kreyer (2003: 170) and Rosenbach (2006: 622-623) – guided the coders' judgments of interchangeability.[1]

3.   A first overview: text frequencies

To provide a first impression of the degree of variation exhibited in the dataset, Figure 1 presents the relative frequency of the *s*-genitive (as a percentage of all interchangeable genitives) across the 10 (sub)corpora studied. There is a good deal of frequency variation: the share of the *s*-genitive ranges from 29.8% in LOB-B to 59.6% in FRED, and while the mean share of *s*-genitive, across all (sub)corpora, is 44.4%, the standard deviation associated with this mean value is a very considerable 10 per cent points.

spoken English	British Dialects (FRED)	59.6%
	colloquial AmE (CSAE)	48.2%
written text types (1990s)	BrE editorials, 90s (F-LOB B)	38.9%
	BrE reportage, 90s (F-LOB A)	50.3%
	AmE editorials, 90s (FROWN B)	49.5%
	AmE reportage, 90s (FROWN A)	55.6%
written text types (1960s)	BrE editorials, 60s (LOB B)	29.8%
	BrE reportage, 60s (LOB A)	41.8%
	AmE editorials, 60s (BROWN B)	30.1%
	AmE reportage, 60s (BROWN A)	39.8%
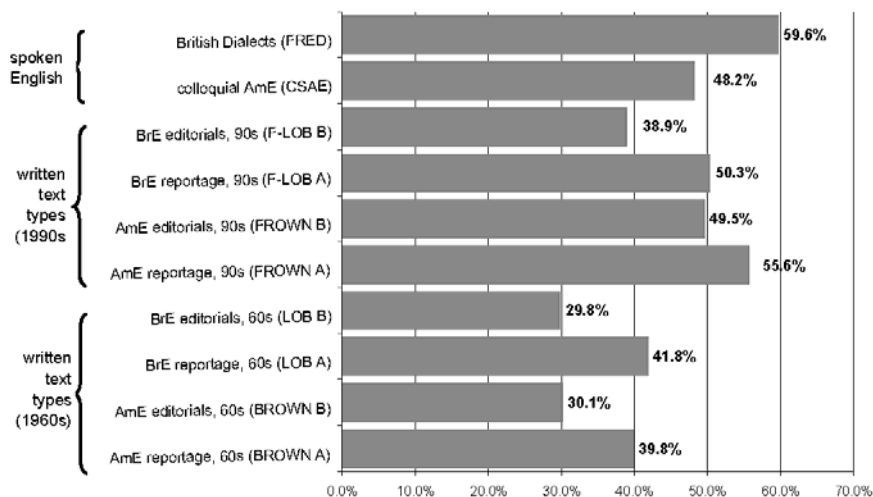
Figure 1. Share of the *s*-genitive among interchangeable genitives across
(sub)corpora

Table 1 elucidates how this variation in text frequency is sensitive to language-external factors. Firstly, as far as sampling time is concerned, the *s*-genitive has become a good deal more frequent in press language in the period between the 1960s and the 1990s, which supports claims in the literature (for instance, Raab-Fischer 1995) that the *s*-genitive is spreading in real time.

Table 1. Mean share of the *s*-genitive among interchangeable genitives according to sampling time, geography, and text type

	mean share of the s-*genitive*
sampling time	
1960s press English	35.4%
1990s press English	48.6%
geography	
American English	44.6%
British English	44.1%
text type	
Spoken	53.9%
press reportage	46.9%
press editorials	37.1%

Secondly, the *s*-genitive is overall a tad more frequent in the American data than in the British data. While the differential is not statistically significant, it nonetheless dovetails with previous claims (cf., for example, Rosenbach 2003: 395-396) that the *s*-genitive is overall more frequent in American English than in British English. Observe, however, that while the *s*-genitive is actually substantially more frequent in American press English than in British press English (mean shares: 43.8% vs. 40.2%), the situation is just the reverse in the spoken data sources, CSAE and FRED (mean shares: 48.2% vs. 59.5%). Notice also that in contemporary press English (as sampled in Frown and F-LOB), the *s*-genitive is significantly more frequent in the American data (mean share: 53.2%) than in the British data (mean share: 45.8%), while the difference between American English and British in the 1960s is marginal (mean shares: 35.0% vs. 36.8%). Thirdly, a fairly neat

text-type continuum emerges: the *s*-genitive is most frequent in spoken data and least frequent in press editorials, while press reportage covers the middle ground. This suggests that press reportage is, compared to press editorials, the more 'colloquial' text type.

The above discussion of overall text frequencies – one-dimensional as they are – has indicated that genitive variation indeed seems to be sensitive to language-external factors. In what will follow, this study will treat text frequencies as epiphenomenal to the probabilistic and cognitive mechanics which underlie the multidimensional system of genitive choice, with a special interest in the role that external factors play in this system.

4.   Conditioning factors in genitive choice

Following the methodology of Szmrecsanyi and Hinrichs (2008), the present study aims to model genitive frequencies as a function of seven major language-internal conditioning factors. These fall into four groups: (i) semantic and pragmatic factors (animacy and thematicity of the possessor), (ii) phonology (i.e. presence of a final sibilant in the possessor), (iii) parsing and processing factors (possessor length, possessum length, and precedence of an identical genitive construction), and (iv) economy (i.e. type-token ratio of a given genitive passage).

4.1. Possessor animacy

Animacy of the possessor NP is commonly claimed to be the chief determinant of genitive choice. Adopting Rosenbach's (2006: 105) animacy hierarchy (human > animal > collective > inanimate) and drawing on Zaenen et al.'s (2004) general coding scheme for animacy, each possessor NP in the dataset was manually annotated according to the following four-way classification: (i) human possessor NPs, as in (3); (ii) animal possessor NPs, as in (4); (iii) collective possessor NPs, as in (5); and (iv) inanimate possessor NPs, as in (6).[2]

(3)   the **emperor**'s family had to call off plans … (Frown A04)

(4)   and he'd pick me up and show me, you know, a little **bird**'s eggs …
      (FRED DEN_001)

(5) Would that the odious discriminatory policy of the **Pentagon** were limited to those two instances. (F-LOB B27)

(6) ... and it was like on the back bumper of the **Honda**, too. (CSAE 0513)

4.2. Thematicity of the possessor NP

According to Osselton (1988), it is the general topic of a text which determines which nouns in that text can take the *s*-genitive. So, while *sound*, *soil*, and *fund* will not normally take the *s*-genitive, "in a book on phonetics, *sound* will get its genitive, in one on farming, *soil* will do so, and in a book on economics you can expect to find *a fund's success*" (Osselton 1988: 143). Assuming, in this spirit, that increased text frequency of a possessor NP would make the *s*-genitive more likely, the *log*-transformed text frequency of the possessor NP's head noun in the respective corpus text (measured in frequency per 2,000 words, which is the standard size of texts in the Brown family) was established for every individual possessor NP in the dataset. The example in (7) will illustrate the basic idea:

(7) The **bill**'s supporters said they still expected Senate approval …
   (Frown A02)

In (7), the possessor NP's head noun is *bill*, and *bill* has a text frequency of 32 occurrences (*log* value: 1.5) in Frown text A02 (which spans about 2,000 words).

4.3. Final sibilants in the possessor NP

A final sibilant in the possessor NP, as in (8), may discourage usage of the s-genitive (cf. Altenberg 1982):

(8) But that is the sad and angry side of **Bush**. (Frown A11)

All possessors in the dataset ending, orthographically, in <s> (as in *Congress*), <z> (as in *jazz*), <ce> (as in *resistance*), <sh> (as in *Bush*), or <tch> (as in *match*) were identified and annotated.[3]

4.4. End weight: possessor and possessum length

The time-honored principle of 'end-weight' (for instance, Behaghel 1909/1910; Wasow 2002) postulates that language users tend to place 'heavier,' more complex constituents after shorter ones, yielding a constituent ordering that might facilitate parsing (see, for example, Hawkins 1994). Hence, if the possessor is heavy, there should be a general preference for the *of*-genitive because it places the possessor last. If the possessum is heavy, a general preference for the *s*-genitive is expected because it places the possessum last. The present study seeks to approximate the weight of genitive constituents by determining their length in graphemic words (see Szmrecsanyi 2004 for an empirical argument that vis-à-vis other measures, counting graphemic words approximates syntactic weight surprisingly well). For illustration, consider (9):

(9)   Latter domain, under **the guidance** of **Chef Tom Yokel**, will specialize
      in steaks, chops, chicken and prime beef as well as Tom's favorite dish,
      stuffed shrimp. (Brown A31)

The possessor phrase in (9) commands three words (*Chef Tom Yokel*) while the possessum spans two words (*the guidance*). Note, though, that if the writer had opted for an *s*-genitive instead, the possessum phrase could not have been determined by an article (**Chef Tom Yokel's the guidance*). Therefore, definite or indefinite articles determining the possessum phrase of an *of*-genitive were not included in the tally (cf. Altenberg 1982: 79-84 for a similar coding procedure). Net possessum length of the possessum phrase in (9) is thus exactly one word (*guidance*).

4.5. Persistence

We now move on to a further processing-related constraint on genitive choice, *viz.* precedence of an identical genitive construction in the preceding textual discourse. We hypothesize that usage of, say, an *s*-genitive in a given genitive slot increases the odds that the speaker/writer will use an *s*-genitive again next time she has a choice (see Szmrecsanyi 2006: 87-107). So, each genitive occurrence in the dataset was annotated according to whether an *s*-genitive had been used last time there was a genitive choice. (10) exemplifies a context where two subsequent interchangeable genitive

contexts (the *continent's river systems* and *the country's Medical Associa-tion*) both exhibit *s*-genitives:

(10)   … the **continent's river systems** are now infected .... In Ecuador, **the country's Medical Association** said 100 people had died of a total of 5,000 cases… (F-LOB A14)

4.6. Lexical density and type-token ratios

Hinrichs and Szmrecsanyi (2007) demonstrate that the *s*-genitive is at-tracted by contexts where informational density is high, i.e. when there is a need to economically code more information in a given textual passage. This is because the *s*-genitive is the more compact and economic coding option (Biber et al. 1999: 99). To check on this factor, Perl scripts estab-lished the type-token ratios of the textual passages (50 words before and 50 words after a given genitive construction) where the genitive occurrences in the dataset were embedded.

5.   Results

5.1. A regression model of genitive choice

We will now draw on *binary logistic regression* (see Pampel 2000) to quantify the combined contribution of the conditioning factors presented above. As a multivariate procedure, logistic regression integrates probabil-istic statements into the description of performance and is applicable "whe-rever a choice can be perceived as having been made in the course of lin-guistic performance" (Sankoff and Labov 1979: 151). In predicting a binary outcome (i.e. a linguistic choice, in the case of the present study whether speakers/writers will choose an *s*-genitive over an *of*-genitive) on the basis of several independent factors (or: predictors), a logistic regres-sion model relies on the following key measures:

–   The magnitude and the direction of the influence of each predictor on the outcome (also known as *factor weights*). This information is pro-

vided by *odds ratios* (ORs), which indicate how the presence or absence of a feature (for categorical factors) or how a one-unit increase in a scalar factor probabilistically influences the odds that some outcome (in our case: choice of an *s*-genitive) will occur. Odds ratios can take values between 0 and ∞: the more the figures exceed 1, the more highly the effect favors a certain outcome; the closer they are to zero (if smaller than 1), the more disfavoring the effect.

– Variability accounted for by (or: explanatory power of) the model as a whole (R^2). The R^2 value can range between 0 and 1 and gauges the proportion of variance in the dependent variable (i.e. in the outcomes) accounted for by all the factors included in the model. Bigger R^2 values mean that more variance is accounted for by the model. The specific R^2 measure which is going to be reported in the present study is the so-called *Nagelkerke R^2*, a pseudo R^2 statistic for logistic regression.

– Predictive efficiency of the model as a whole. The percentage of correctly predicted cases (*% correct*) vis-à-vis the baseline prediction (*% baseline*) indicates how accurate the model is in predicting actual outcomes. The higher this percentage, the better the model.

Rather than fitting a one-size-fits-all regression model on the entire dataset and modeling the effect of external factors via interaction terms, the present investigation fits 10 independent regression models – one for each of the (sub)corpora under analysis – on the language-internal factors discussed in section 4 above.[4] The results are provided in Table 2. Predictive efficiency of the models is satisfactory: on the basis of the conditioning factors considered, the models predict between 70.4% (CSAE) and 88.8% (FRED) of the genitive outcomes accurately. Variance explained (R^2) ranges between .34 (LOB-B) and .68 (FRED), which is another way of saying that we can account for between 34% and 68% of the observable variability in the (sub)corpora under analysis – the remainder of the variability may be due to free variation, or to other conditioning factors not considered in the present study. In all, the system of genitive choice sketched in Table 2 works best for the very traditional dialect speech sampled in FRED, and least well (though still somewhat satisfactorily) for 1960s British English press editorials, as sampled in LOB-B. There is, moreover, a tendency for those models on spoken data to have a better fit than models on written data (mean R^2 spoken data: .56, mean R^2 spoken data: .45), which may suggest that in written data, other factors not considered here (stylistics, prescriptivism, etc.) might have more weight than in spoken data.

Table 2. Odds ratios (ORs) in logistic regression. Predicted odds are for the *s*-genitive. Significant ORs ($p < .05$) are in bold

		CSAE	FRED	Brown-A	Brown-B	LOB-A	LOB-B	Frown-A	Frown-B	F-LOB-A	F-LOB-B
possessor animacy (default category: inanimate)	human	**8.08**	**69.66**	**8.53**	**13.00**	**11.01**	**18.40**	**7.25**	**6.76**	**13.84**	**15.36**
	animal	**30.94**	**17.75**	.00	.00	99	6.56	.00	1.55	99	.00
	collective	**3.94**	**2.77**	**3.40**	**3.35**	**3.16**	**3.91**	**3.63**	**2.69**	**4.86**	**5.53**
thematicity of possessor		.90	.99	1.20	**1.50**	1.20	1.25	**2.14**	**1.95**	**1.82**	**1.29**
final sibilant in possessor		.21	.36	.24	.25	.50	.54	.22	.22	.30	.27
possessor length		.52	.42	.38	.44	.42	.63	.41	.55	.40	.42
possessum length		1.00	.96	1.19	1.45	1.16	.95	**1.45**	**1.57**	**1.54**	**1.97**
persistence		**3.53**	**1.87**	1.66	1.51	1.44	**2.11**	1.31	1.34	1.21	1.38
type-token ratio		.90	.95	**2.42**	**1.77**	**2.36**	**2.55**	**1.68**	**2.10**	**2.23**	**1.58**
N		332	1,818	1,329	804	1,241	707	1,244	816	1,138	736
% baseline		52.0	59.6	60.0	70.1	58.4	70.7	55.5	50.6	50.6	61.1
% correct		70.4	88.8	75.7	80.8	75.2	78.8	77.7	76.1	80.5	79.2
Nagelkerke R^2		.43	.68	.45	.46	.42	.34	.48	.42	.54	.49

Let us next discuss individual factor groups and their effect on genitive choice. As for semantic and pragmatic factors, consider animacy. The models reported in Table 2 take inanimate possessors (*the Honda*, *a rock*, etc.) as the default category and quantify the effect that human/animal/collective possessors have on the odds that an *s*-genitive will be chosen. The effect of human and collective possessors is statistically significant throughout, while animal possessors are significant in the spoken corpora only (the simple reason for this being that animal possessors are a very rare species in press material). The factor also has the theoretically expected effect direction: as a generalization, the more animate a possessor is, the greater the odds that an *s*-genitive will be chosen. Take, for instance, Brown-A: if the possessor is animate (e.g. *the emperor*, *John*) instead of inanimate (e.g. *the Honda*, *a rock*), the odds that an *s*-genitive will be chosen increase by a factor of 8.53. If the possessor is a collective noun (e.g. *the Pentagon*, *the police*), the odds for an *s*-genitive increase by a factor of 3.40. Notice now that there is a general tendency for human possessors to attract s-genitives more strongly in the British data (mean OR: 25.65) than in the American data (mean OR: 8.72), suggesting that the *s*-genitive is cognitively more strongly associated with human possessors in British English than in American English. FRED is an extreme case: the huge odds ratio of 69.66 associated with human possessors indicates that in traditional British dialects, human possessors – for all intents and purposes – categorically trigger the *s*-genitive. This is unlikely to be due to, e.g., the text type (interviews) sampled in FRED. Instead, what we are seeing here is probably an older system of genitive choice, given that informants in FRED are elderly people and that many of the traditional dialects sampled in the corpus are rather conservative. Notice here that this line of reasoning does not contradict the fact that the *s*-genitive is becoming more frequent in Present-Day English – the contemporary expansion of the *s*-genitive in press English is actually quite unrelated to the animacy constraint.

As detailed above, the literature suggests that increased thematicity of the possessor – operationalized as the possessor head noun's *log* text frequency in a given corpus text – makes the *s*-genitive more likely. In the written data sources, this hypothesis is indeed borne out: for every one-unit increase in a possessor head's *log* text frequency (to illustrate, this would correspond to a frequency differential of, very roughly, 3 occurrences per corpus text instead of 1 occurrence per corpus text), the odds for the *s*-genitive increase by a factor of between 1.20 (Brown-A, LOB-A) and 2.14 (Frown-A). Overall, the factor appears to be somewhat more powerful in

the written American data (mean OR: 1.70) than in the British data (mean OR: 1.39). It is also stronger in 1990s texts (mean OR: 1.80) than in 1960s texts (mean OR: 1.29). By contrast, the factor is not even selected as significant in the spoken corpora (CSAE and FRED). In other words, possessor thematicity is characteristic of written, not spoken, language.

Turning to phonology, a final sibilant in the possessor significantly and reliably discourages usage of the *s*-genitive, as expected: the presence of a final sibilant decreases the odds for an *s*-genitive by between 46% (LOB-B) and 79% (CSAE). There is hardly any difference between the written (mean OR: .32) and the spoken data sources (mean OR: .29), though interestingly the constraint has become significantly (cf. Hinrichs and Szmrecsanyi 2007) more influential over time in press language (mean OR 1960s: .28, mean OR 1990s: .25). The somewhat curious fact that a phonological constraint should become more influential in press language (a written genre) over time advertises itself to be interpreted in terms of a "colloquialization of the norms of written English" (Leech and Smith 2006; Hundt and Mair 1999).

What about factors relating to parsing and processing? As hypothesized, longer possessor NPs significantly and consistently disfavor the *s*-genitive (because this coding option places the possessor second): for every additional word in the possessor NP, the odds for an *s*-genitive decrease by between 62% (Brown-A) and 37% (LOB-B), an effect which, among the written data sources, is stronger in press reportage material (mean OR: .40) than in press editorials material (mean OR: .51). Conversely, longer possessum NPs significantly attract the *s*-genitive in six of the ten data sources studied: thus, for every additional word in the possessum NP, the odds for an *s*-genitive increase by between 19% (Brown-A) and 97% (F-LOB-B). In this connection it should be noted that possessum length does not seem to be important in the spoken data sources, which is another way of saying that the factor is a characteristic of the written, not spoken, English system of genitive choice.

The factor 'persistence' is significant in six of the ten (sub)corpora studied (it is not significant in 1990s press English), and has the theoretically expected sign throughout: among the data sources where the factor is significant, precedence of an *s*-genitive in the ongoing discourse increases the odds for another, subsequent *s*-genitive by a factor of between 1.44 (LOB-A) and 3.53 (CSAE). In all, it is fairly evident that persistence effects are more important in the spoken data sources than in the written data sources, which hardly comes as a surprise given the effect's deep rootedness in the

nature of online processing constraints (on this point, cf. Szmrecsanyi 2005a,b, 2006).

We finally move on to the economy-motivated factor in the variable portfolio, *viz*. lexical density as approximated by the type-token ratio of a given genitive passage. Recall that we assumed that speakers/writers would resort to the more economical *s*-genitive in contexts characterized by high type-token ratios and thus high lexical (or: informational) density. For writers (though not for speakers), this hypothesis is borne out: for every 10-word increase in a given genitive context's type-token ratio (if, say, such a context contains 70 different types, instead of just 60), the odds for an *s*-genitive increase by a factor of between 1.58 (F-LOB-B) and 2.55 (LOB-B). Because the predictor is not even selected as significant in the spoken data sources, the sort of economy implicit in the nature of the predictor appears not to be important in spoken language.

By way of an interim summary, the most important finding of this portion of the analysis is that the grammar of genitive choice is *qualitatively* (that is, in terms of the effect direction of the factors studied) very similar in all of the ten (sub)corpora under investigation. Where significant, more animate and thematic possessors, longer possessum phrases, precedence of an *s*-genitive, and higher type-token ratios all attract the *s*-genitive. Final sibilants and long possessor phrases, in turn, attract the *of*-genitive. At the same time, we have seen that the *magnitude* of the effect of individual predictors may vary, statistically, as a function of a number of language-external factors – time, geography, and text type. In an attempt to see the wood for the trees, it should be worthwhile to invoke this quantitative variance to establish aggregate similarities (and dissimilarities) between the cognitive and probabilistic grammars of genitive choice. It is to this task that I next turn.

5.2. Aggregate similarities between genitive choice systems

Thus far, we have sought to characterize the cognitive and probabilistic grammar of genitive choice in English on the basis of a complex system of conditioning factors, yielding ten sets of nine discrete odds ratios – one for each data source under analysis – which characterize this system. Note, now, that fine-grained and instructive as the analysis of conditioning factors may be, its multidimensional nature makes it rather difficult to spot overarching tendencies and patterns relying merely on one's eyeballs. This

is why we will now abandon our earlier focus on individual factors and their probabilistic weights, turning instead to two non-parametric statistical analysis and visualization techniques (*cluster analysis* and *multidimensional scaling*) to uncover the 'big' picture of genitive variation in time, geography, and across text types.
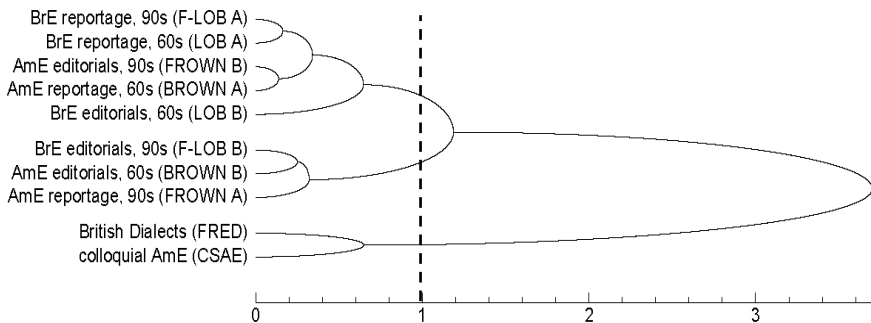
Figure 2. Dendrogram derived from hierarchical agglomerative cluster analysis (cluster algorithm: Ward's method) of the *log*-transformed 10 × 9 odds ratio matrix in Table 2

Cluster analysis is a cover term for a set of techniques designed to objectively group a given number of cases (in this study, probabilistic grammars of genitive choice) into a smaller number of discrete and meaningful clusters on the basis of some sort of similarity – in our case, similarities between probabilistic factor weights – in order to establish higher-order patterns in an objective way (for an introduction to the technique from the social scientist's perspective, see Aldenfelder and Blashfield 1984). Data clustering can be visually represented using tree diagrams, also known as *dendrograms*, which work in essentially the same way as family trees. The dendrogram deriving from this study's dataset (more specifically, from the probabilistic factor weights in Table 2) can be seen in Figure 2. [5] In this dendrogram, the first and most basic split occurs between the written and the spoken (sub)corpora under investigation. Further down the road, the written cluster regroups into two subclusters, yielding a three-cluster solution at a (statistically comparatively robust) cophenetic distance of 1.0, as indicated by the dotted vertical line in Figure 2. The first of the two written subclusters contains the British press reportage subcorpora (LOB-A, F-LOB-A), the 1960s British press editorials subcorpus (LOB-B), and the two American subcorpora (Frown-B, Brown-A). The second of the two written

subclusters is more homogeneous, containing 1990s British editorials (F-LOB-B) as well as the remainder of the American material (Brown-B, Frown-A). In all, Figure 2 makes amply clear that the most fundamental split – as indicated by the distance from the leaves to the encompassing node – in the dataset occurs between the written and the spoken material, which testifies to the paramount importance of the written-spoken distinction for the exact quantitative shape of a given system of genitive choice. This distinction overrides all other language-external factors.
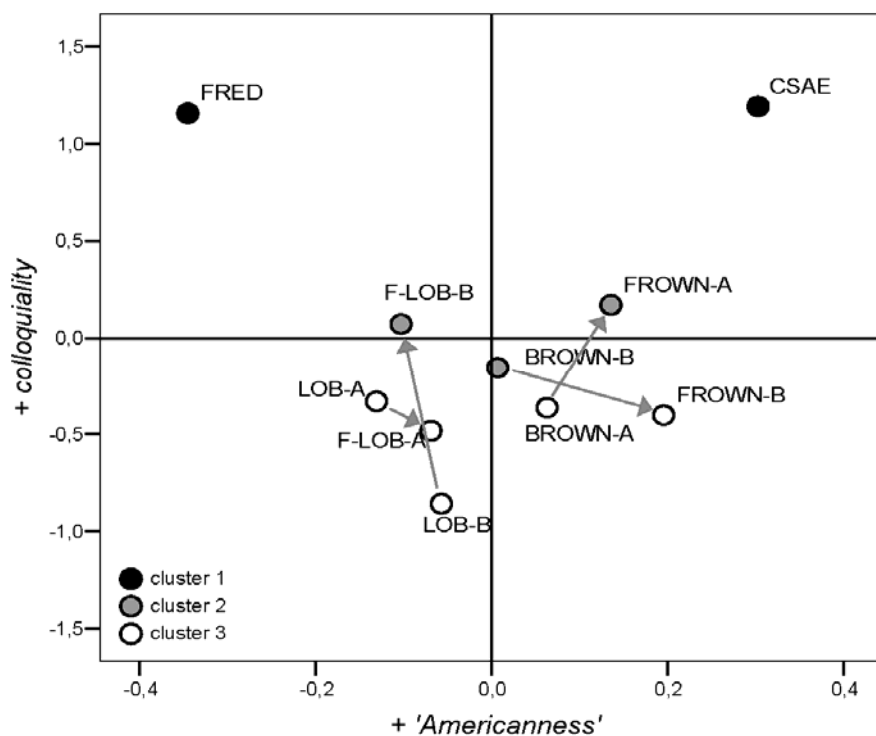
Figure 3. MDS visualization of the *log*-transformed 10 × 9 odds ratio matrix in Table 1. Group memberships derive from hierarchical agglomerative cluster analysis (cf. Figure 2). Arrows indicate drifts in real time

The dendrogram in Figure 2 has provided us with a first impression of the similarities and dissimilarities between genitive choice systems as exhibited in our dataset. For the remainder of this section, we will rely on *multidimensional scaling* to visualize the hidden structure of genitive variation in

time, space, and across text types (for an introduction to multidimensional scaling, see Kruskal and Wish 1978). This means that we will scale down the original nine dimensions (i.e. probabilistic factor weights) by which every genitive choice system in our dataset is characterized to two dimensions, an exercise which will make it possible to visualize the aggregate (dis-)similarities between these systems in two-dimensional maps. The advantage of such perceptual maps is that these can be interpreted fairly intuitively: much as with geographic maps, the further two points are apart, the more dissimilar (in geographic terms, distant) they are. If two pairs of points are equally close or distant, the pairs of genitive choice systems they represent are equally (dis-)similar.[6] The resulting visualization is given in Figure 3; also shown in this figure are cluster memberships as derived from hierarchical agglomerative clustering (see Figure 2).

We observe, first and foremost, that the relative distance between the spoken material in FRED and the CSAE (cluster 1) and the written macro cluster (clusters 2 and 3) is considerable. So, in a bird's eye perspective, the written (sub)corpora clearly form a genre of their own, which is different from the spoken material. What is happening within the written text types, though? To begin to address this question, consider the position of the data points relative to the vertical axis: high values (as in cluster 1) are associated with spoken material, so the vertical axis may be considered indicative of increasing levels of orality, i.e. *colloquiality*. Assuming that this interpretation is correct, the material in cluster 3 is least colloquial, while the material in cluster 2 covers the middle ground. It turns out, therefore, that cluster analysis has grouped the material in the dataset according to increasing levels of colloquiality as, once again, the most important external parameter working on genitive choice systems. What is the interpretation of the horizontal axis? Observe that all data sources yielding negative scores on the horizontal axis sample British material, while all the data sources yielding positive values comprise American material. The horizontal axis may thus be considered being indicative of increased '*Americanness*'.

The *colloquiality* vs. '*Americanness*' dimensions underlying the plot in Figure 3 yield an additional four-way classification of the material in our dataset: the upper left-hand quadrant in Figure 3 is the *colloquial/British* quadrant, the upper right-hand quadrant is the *colloquial/American* quadrant, the lower right-hand quadrant is the *written/American* quadrant, and the lower left-hand quadrant is the *written/British* quadrant. Having so at once taken care of the external parameters 'text type' (spoken vs. written)

and 'geography', we will now go on to a discussion of drifts, among the written material, in real time. Recall from the literature review that Hinrichs and Szmrecsanyi (2007: 469) have shown that in written English in particular, it is primarily a process of 'economization' that drives the spread of the *s*-genitive in real time. Szmrecsanyi and Hinrichs (2008) – not differentiating between the written genres (press editorials vs. press reportage) that are subject to differentiation in the present study – likewise suggest that press language *as such* cannot be said to have substantially colloquialized. In the light of the present study's more fine-grained distinction between press reportage and press editorials, and on the basis of Figure 3 (consider the arrows indicating diachronic drifts), these claims can be restated more precisely in the following way:

–   *British press reportag*e (LOB-A → F-LOB-A) exhibits a modest drift towards less colloquiality ('de-colloquialization') as well as modest Americanization;

–   *British press editorials* (LOB-B → F-LOB-B) attest a considerable drift towards more colloquiality ('colloquialization') as well as modest shift away from the American sector of the diagram;

–   *American press reportage* (Brown-A → Frown-A) shows a medium-scale drift towards more colloquiality ('colloquialization') and slight Americanization (to the extent, of course, that a *per se* American genre can become even more American);

–   *American press editorials* (Brown-B → Frown-B) are characterized by a modest shift towards less colloquiality ('de-colloquialization') and medium-scale Americanization (cf. the caveat above).

This exercise in drift tracing has suggested that curiously – as far as the direction of the drifts (and not the respective endpoints) are concerned – British press reportage aligns with American press editorials, and British press editorials somewhat align with American press reportage. In sum, the data reveal that while consistent with extant literature there is no such thing as a robust *overall* pattern of colloquialization or Americanization in press English, the two processes are arguably still somewhat involved in diachronic drift, depending on text type and geographic variety. The mediating factor that very likely accounts for this interpretatorial twilight is *economi-*

zation, *viz*. the differential importance, depending on text type, of the "tendency to brevity" (Dahl 1971: 172) and of the need to save paper space by opting for more compact coding options (such as the *s*-genitive) instead of more explicit coding options (such as the *of*-genitive). Because such pressures are arguably more acute in press reportage than in editorials, we see differential drift directions (a more detailed discussion of this issue is provided in Szmrecsanyi and Hinrichs 2008).

6.   Concluding remarks

The foregoing analysis leads to two principal conclusions about the alternation between the *s*-genitive and the *of*-genitive in English in a cognitive sociolinguistics perspective. For one thing, we have seen that while there is a good deal of variation in text frequencies, the probabilistic grammar underlying the system of genitive choice is fundamentally the same across sampling times, geographic varieties of English, and text types: animate possessors are cognitively associated with the *s*-genitive, long possessor NPs trigger the *of*-genitive, and so on. On the other hand, however, the magnitude of the effect that individual conditioning factors may have on genitive choice can vary substantially across different data sources, and this statistical variance is demonstrably mediated by language-external factors. By aggregating individual factor weights to an aggregate measure of distance between genitive choice systems and by subsequently partitioning and visualizing the resulting variance, this study has sought to demonstrate that the most important language-external factor working on the English genitive alternation is the written/spoken text-type distinction, and that the real-time drift of written genitive choice systems – depending on their exact genre and on whether they are British or American – may be differentially impacted by cultural phenomena such as colloquialization, Americanization, or economization. On more methodological grounds, this study highlights the fact that by exploring how language-external and cultural factors leave their mark on the quantitative footprint of probabilistic grammars, and thus on the cognitive factors that motivate linguistic choices, we can learn a lot about how language variation is more patterned and predictable than one might perhaps think. In exactly this spirit, further study may wish to continue this line of inquiry to explore, e.g., how genuinely sociological variables such as age, gender, and social class interact with probabilistic grammars.

Notes

1. As for interrater reliability, parallel annotation of a set of $N = 202$ genitives by two trained coders yielded (i) a simple agreement rate of 86% and a "good" (cf. Orwin 1994: 152) Cohen's κ value of .69 for *s*-genitives, and (ii) a simple agreement rate of 89% and an "excellent" Cohen's κ value of .78 for *of*-genitives. Hinrichs and Szmrecsanyi (2007: section 3) provide more detail.

2. Interrater reliability of animacy coding was satisfactory: parallel coding of a random subset of $N = 199$ genitive possessors by two trained coders yielded a simple agreement rate of ca. 86% and an "excellent" (cf. Orwin 1994: 152) Cohen's κ value of .79. Hinrichs and Szmrecsanyi (2007: section 5.1.1) provide more detail.

3. Possessors ending in <dge> (as in *judge*) are so rare that they were excluded from analysis.

4. Note that this is mainly for expository purposes – interaction terms can be notoriously hard to interpret. Also notice that the analysis techniques drawn on in Section 5.2. (cluster analysis and multidimensional scaling) will draw on the discrete odds ratio vectors presented in Table 2. See Hinrichs and Szmrecsanyi (2007) for a uniform model of genitive choice in the Brown family of corpora that models the effect of language-external factors as interaction terms. I should also like to point out that in the present study's dataset, there are no statistically significant and/or substantially interpretable interaction effects *between* the language-internal factors considered here (say, between animacy and thematicity).

5. Technically, the set of 10×9 odds ratios in Table 2 was first *log*-transformed (in order to alleviate the effect of outliers) and then converted into a distance matrix using Euclidean distance as an interval measure. On the basis of this distance matrix, a hierarchical agglomerative clustering algorithm (specifically, Ward's Minimum Variance method) subsequently partitioned the (sub)corpora in the dataset into clusters. Note that because simple clustering can be unstable (see, for instance, Nerbonne et al. 2007), the robustness of the dendrogram in Figure 2 was assessed by also running three other common clustering algorithms – Weighted Average (WPGMA), Group Average (UPGMA), and Complete Link – on the dataset. Since the exact same dendrogram as reported in Figure 2 also emerged in two of the three additional runs (with only the Complete Link algorithm yielding a slightly different clustering outcome), the dendrogram in Figure 2 can be considered fairly reliable.

6. The scaling procedure was conducted using the Proxscal algorithm implemented in SPSS, on the basis of the same distance matrix (derived from Euclidean distances in the *log*-transformed set of 10×9 odds ratios) used as input to the cluster analysis (see previous footnote). The resulting two-dimensional scaling solution yields a normalized raw stress value of .0012, a dispersion-

accounted-for value of .99, and a Tucker's coefficient of congruence value of .99.

References

Aldenfelder, Mark, and Roger Blashfield
1984   *Cluster Analysis*. Newbury Park/London/New Delhi: Sage Publications.

Altenberg, Bengt
1982   *The Genitive v. the Of-Construction. A Study of Syntactic Variation in 17th Century English*. Malmö: CWK Gleerup.

Behaghel, Otto
1909/10   Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen* 25: 110-142.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan
1999   *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Dahl, Lisa
1971   The *s*-genitive with non-personal nouns in modern English journalistic style. *Neuphilologische Mitteilungen* 72: 140-172.

Du Bois, John W., Wallace L. Chafe, Charles Meyer, and Sandra A. Thompson
2000   *Santa Barbara Corpus of Spoken American English*, Part 1. Philadelphia: Linguistic Data Consortium.

Du Bois, John W., Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, and Nii Martey
2003   *Santa Barbara Corpus of Spoken American English*, Part 2. Philadelphia: Linguistic Data Consortium.

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

Hernández, Nuria
2006   *User's Guide to FRED*. Available online at http://www.freidok.uni-freiburg.de/volltexte/2489 (last accessed: 3/13/2008). Freiburg: English Dialects Research Group.

Hinrichs, Lars, Birgit Waibel, and Nicholas Smith
2007   The POS-tagged, postedited F-LOB and Frown corpora: a manual, including pointers for successful use. Freiburg: Department of English, University of Freiburg.

Hinrichs, Lars, and Benedikt Szmrecsanyi
  2007   Recent changes in the function and frequency of standard English genitive constructions: a multivariate analysis of tagged corpora. *English Language and Linguistics* 11 (3): 437-474.

Hundt, Marianne, and Christian Mair
  1999   'Agile' and 'uptight' genres: the corpus-based approach to language change in progress. *International Journal of Corpus Linguistics* 4: 221-242.

Jucker, Andreas
  1993   The genitive versus the *of*-construction in newspaper language. In *The Noun Phrase in English. Its Structure and Variability*, Andreas Jucker (ed.), 121-136. Heidelberg: Carl Winter

Kreyer, Rolf
  2003   Genitive and *of*-construction in modern written English. Processability and human involvement. *International Journal of Corpus Linguistics* 8 (2): 169-207.

Kruskal, Joseph, and Myron Wish
  1978   *Multidimensional Scaling.* Newbury Park/London/New Delhi: Sage Publications.

Leech, Geoffrey, and Nicholas Smith
  2005   Extending the possibilities of corpus-based research on English in the twentieth century: a prequel to LOB and F-LOB. *ICAME Journal* 29: 83-98.
  2006   Recent grammatical change in written English 1961-1992: some preliminary findings of a comparison of American with British English. In *The Changing Face of Corpus Linguistics*, Antoinette Renouf, and Andrew Kehoe (eds.), 185-204. Amsterdam/New York: Rodopi.

Mair, Christian
  2006   Inflected genitives are spreading in present-day English, but not necessarily to inanimate nouns. In *Corpora and the History of English: Festschrift für Manfred Markus*, Christian Mair, and Reinhard Heuberger (eds.), 243-256. Heidelberg: Winter.

Nerbonne, John, Peter Kleiweg, Franz Manni, and Peter Heeringa
  2007   Projecting dialect differences to geography: bootstrap clustering vs. noisy clustering. In *Data Analysis, Machine Learning, and Applications. Proc. of the 31st Annual Meeting of the German Classification Society*, Christine Preisach, Lars Schmidt-Thieme, Hans Burkhardt, and Reinhold Decker (eds.). Berlin: Springer.

Orwin, Robert
    1994   Evaluating Coding Decisions. In *The Handbook of Research Synthe-sis*, Harris Cooper, and Larry Hedges (eds.), 139-162. New York: Russell Sage Foundation.
Osselton, Noel
    1988   Thematic Genitives. In *An Historic Tongue: Studies in English Lin-guistics in Memory of Barbara Strang*, Graham Nixon, and John Honey (eds.). London: Routledge.
Pampel, Fred
    2000   *Logistic Regression. A Primer*. Thousand Oaks: Sage Publications.
Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik
    1985   *A Comprehensive Grammar of the English Language*. London, New York: Longman.
Raab-Fischer, Roswitha
    1995   Löst der Genitiv die *of*-Phrase ab? Eine korpusgestützte Studie zum Sprachwandel im heutigen Englisch. *Zeitschrift für Anglistik und Amerikanistik* 43(2): 123-132.
Rosenbach, Anette
    2002   *Genitive Variation in English: Conceptual Factors in Synchronic and Diachronic Studies*. Berlin/New York: Mouton de Gruyter.
    2003   Aspects of iconicity and economy in the choice between the *s*-genitive and the *of*-genitive in English. In *Determinants of Gram-matical Variation in English*, Günter Rohdenburg, and Britta Mon-dorf (eds.), 379-412. Berlin/ New York: Mouton de Gruyter.
    2006   Descriptive genitives in English: a case study on constructional gra-dience. *English Language and Linguistics* 10 (1): 77-118.
Sankoff, David, and William Labov
    1979   On the use of variable rules. *Language in Society* 8: 189-222.
Szmrecsanyi, Benedikt
    2004   On operationalizing syntactic complexity. In *Le poids des mots. Proceedings of the 7th International Conference on Textual Data Statistical Analysis. Louvain-la-Neuve, March 10-12, 2004*, Pur-nelle, Gérard, Cédrick Fairon, and Anne Dister (eds.), 1032-1039. Louvain-la-Neuve: Presses universitaires de Louvain.
    2005a   Language users as creatures of habit: a corpus-linguistic analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory* 1 (1): 113-150:
    2005b   Never change a winning chunk. *Recherches Anglaises et Nord-Américaines* 38: 21-34.
    2006   *Morphosyntactic Persistence in Spoken English. A Corpus Study at the Intersection of Variationist Sociolinguistics, Psycholinguistics, and Discourse Analysis.* Berlin/New York: Mouton de Gruyter.

Szmrecsanyi, Benedikt, and Nuria Hernández
    2007   *Manual of Information to accompany the Freiburg Corpus of English Dialects Sampler ("FRED-S")* Available online at http://www.freidok.uni-freiburg.de/volltexte/2859/ (last accessed: 3/13/2008). Freiburg: English Dialects Research Group.
Szmrecsanyi, Benedikt, and Lars Hinrichs
    2008   Probabilistic determinants of genitive variation in spoken and written English: a multivariate comparison across time, space, and genres. In *The Dynamics of Linguistic Variation: Corpus Evidence on English Past and Present*, Terttu Nevalainen, Irma Taavitsainen, Päivi Pahta, and Minna Korhonen (eds.), 291-309. Amsterdam: Benjamins.
Wasow, Thomas
    2002   *Postverbal behavior*. Stanford, CA: CSLI Publications.
Zaenen, Annie, Jean Carlette, Gregory Garretson, Joan Bresnan, Andrew Koontz-Garboden, Tatiana Nikitina, M. Catherine O'Connor, and Thomas Wasow
    2004   Animacy encoding in English: why and how. In *Proceedings of the 2004 ACL workshop on discourse annotation, Barcelona, July 2004,* Donna Byron and Bonnie Webber (eds.), 118-125.

(Not) acquiring grammatical gender in two varieties of Dutch

Gunther De Vogelaer

Abstract

This paper discusses regional differences in the way the gender system is acquired in two varieties of Dutch, one spoken in the Netherlands (province of Overijssel) and one in the Belgian province of East Flanders. The traditional grammatical three-gender system plays a much more important role in the Belgian child data than in the data from the Netherlands. This is most clearly visible through the fact that East Flemish children show attestations of feminine gender for non-animate count nouns and mass nouns, whereas the children from Overijssel only use feminine pronouns to refer to female humans or animates. Thus, the way in which the gender system is acquired mirrors the rule system that is used by adults in either variety of Dutch (cf. Mills 1986 on German and English). Essentially, however, the pronominal gender system of both the Overijssel and the East Flemish children must be characterized as predominantly semantic. From this, a tentative conclusion can be drawn about the future of Dutch: given Nesset's (2006) Core Semantic Override Principle, it is likely that northern and southern Dutch pronominal gender will ultimately converge in a system of semantic agreement.

Keywords: language acquisition, grammatical gender, semantic gender, geographical variation, pronouns, Dutch

1. Introduction[1]

A tacit assumption underlying much work in first language acquisition is that, in dealing with the acquisition of a certain (standard) language, all or most children behave alike, even when they have a different geographical or social background. Hence it is not very common that details are provided as to the regional or social background of the investigated children, apart perhaps from factors involving bilingualism. Given that even standard languages display significant amounts of regional, social or other types of

variation, however, this may very well be a problematic assumption. In this paper it is shown that there are indeed cases in which the acquisition of an aspect of grammar differs between varieties of one language. More specifically, regional differences are discussed in the way the gender system is acquired in two varieties of Dutch, one spoken in the Netherlands and one in the Dutch-speaking north of Belgium. That the relevant varieties are spoken in different countries does not entail that the variation can be described as resulting from the existence of different national varieties of Dutch. Since both in the Netherlands and in Belgium there is significant variation in the way grammatical gender is used by adults, variation is equally likely to occur within the boundaries of each of the two countries under investigation.

The paper is organized as follows: in section 2, the Dutch gender system is described, both in synchronic and diachronic terms, and some of the literature on the acquisition of gender in Dutch, German and English is discussed. Section 3 and 4 provide an overview of the results of a survey on pronominal gender in a Dutch and a Belgian variety of Dutch, respectively. In section 5 the data for the two varieties of Dutch under discussion are compared with each other and with the literature on the acquisition of gender in German and English. From this comparison a number of predictions can be drawn about the future of the Dutch gender system. Section 6 concludes this paper.

2.   Gender systems and language acquisition: the Dutch case

2.1. Dutch gender in transition

Recent centuries have seen dramatic changes in the Dutch gender system (see Geerts 1966 for a description). Historically, Dutch, like German, distinguished three genders: masculine, feminine and neuter. Gender was marked in two ways: first, adnominal elements such as articles and adjectives agree in gender with the head noun of the noun phrase ('adnominal gender'). Second, personal pronouns take over the gender of the antecedent noun to which they refer ('pronominal gender'). As for the way gender is assigned, there is no apparent semantic motivation for gender in Dutch, although the gender for nouns referring to humans in general corresponds

to the natural gender of the referents (the most important exceptions are the neuters *meisje* 'girl' and *kind* 'child'). Hence the system is described as a so-called 'grammatical' gender system (cf. the terminology in Corbett 1991).

In modern Standard Dutch, quite different developments are observed in adnominal and pronominal gender. In the adnominal domain, the traditional masculine and feminine genders have collapsed, giving rise to a two-gender system in which common gender is distinguished from the neuter. This innovative two-gender system determines which article will be used (cf. the terms *de*-words for common and *het*-words for neuter nouns), how adjectives are inflected, and which demonstrative, interrogative and relative pronouns are used. Table 1 shows the definite and indefinite articles. Although the number of genders has changed, the system in Table 1 remains a grammatical one, i.e. there is still no underlying semantic motivation determining the gender of the nouns.

Table 1. Adnominal gender in Dutch

	present-day Standard Dutch		three-gender system in southern dialects	
	indefinite article	*definite article*	*indefinite article*	*definite article*
masculine:	een	de	ne(n)	de(n)
feminine:	een	de	een	de
neuter:	een	het	ee(n)	het

The traditional three-gender system has not disappeared completely, however; it is still found in substandard varieties and dialects, including most dialects spoken in Belgium (see De Schutter et al. 2005 for dialect maps on adnominal gender). Table 1 includes the article system found in Brabantic and East Flemish dialects, which still distinguishes three genders. There are two important differences between this dialectal system and the Standard Dutch one. First, the masculine indefinite article is *ne(n)* rather than *een*; and second, masculine nouns trigger an inflectional *-n* on both definite and indefinite articles, which is dropped when it is not followed by a vowel, /h/, /b/, /d/ or /t/ (Taeldeman 1980). The final /n/ in the neuter indefinite article *een* is dropped under the same conditions. Similar differences between the

dialectal system and the Standard Dutch one are found on other gender agreement targets (e.g. demonstratives, adjectives,... see Taeldeman 1980 and Plevoets, Speelman and Geeraerts, *to appear* for a description).

According to Dutch reference grammars such as Haeseryn et al. (2002), pronominal gender has changed accordingly, i.e. the use of personal pronouns such as the subject forms *hij* 'he', *ze* 'she' and *het* 'it', the object forms *hem* 'him', *haar* 'her' and *het* 'it' and the possessives *zijn* 'his/its' and *haar* 'her' to refer to antecedent nouns.[2] Especially the north of the Dutch language area has seen a tendency to use etymologically masculine pronouns in reference to both masculine and feminine nouns. The most visible effect of such a system is the near-disappearance of the feminine pronoun *ze* 'she', the use of which is confined to reference to female humans. Hence the system can be characterized as a grammatical two-gender system. According to Audring (2006), however, some varieties are undergoing a more radical change in their pronominal usage, in that the grammatical system is being replaced by an innovative, semantic system. The main parameter in this system is individuation: a high degree of individuation is associated with etymologically masculine pronouns, a low degree of individuation with neuter pronouns. The different behavior of count nouns and mass nouns illustrates this system: countable nouns tend to trigger masculine pronouns such as the weak form *'m* 'him' in (1a); mass nouns are increasingly referred to with neuter pronouns such as *'t* 'it' in (1b). A similar system is found in some regional varieties of English (Siemund 2002, 2008). Resemantisation is reported only in personal pronouns and in relative pronouns, and does not appear to affect the adnominal system.

(1)   Pronominal gender in northern vs. southern varieties (northern examples from Audring 2006:95-96

   a. [+count]:about dat boek 'that book' (neuter):
      North:'Dan moet 'k '**m** ook niet gaan inleveren' (semantic gender)'Then I shouldn't return **him** yet'
      South:'Dan moet 'k '**t** ook niet gaan inleveren' (grammatical gender)
      'Then I shouldn't return **it** yet'

   b. [-count]: about olijfolie 'olive oil' (common/feminine):
      North: '... hoe '**t** geconserveerd wordt' (semantic gender)
      '... how **it** is preserved'
      South: '... hoe **ze** geconserveerd wordt' (grammatical gender)
      '... how **she** is preserved'

As for pronominal gender, both grammatical systems, i.e. the three-gender system and the innovative dyadic one, are considered to be Standard Dutch (cf. Haeseryn et al. 2002, §3.3.3). But apart from southern speakers living in the dialect areas of Brabant or Limburg (see Hoppenbrouwers 1983 and Vousten 1995:73, respectively), speakers of Dutch in the Netherlands no longer use the traditional three-gender system. For Belgium, a question-naire study on gender (Geeraerts 1992:75) reveals that the majority of ref-erences to grammatically feminine nouns in Belgian Standard Dutch are in accordance with grammatical gender: for most items, between 60 and 100% feminine answers are observed. Slightly over 20% masculine pronouns are found; the neuter pronoun *het* 'it' is only marginally observed.

The semantic system described by Audring (2006) is not (yet?) endorsed in normative sources (although it is not described as non-standard usage either), and in fact relatively little is known about the extent to which this system has diffused. Audring (2006:111-112) bases her description on an analysis of the Corpus of Spoken Dutch. She estimates that in informal Dutch as spoken in the north, 71% of pronominal references are semanti-cally motivated. The question whether this system is as pervasively found in more formal registers, in written Dutch, or in other regions remains by and large unaddressed. De Vogelaer (2009:77) provides data for the di-alects spoken in the Belgian provinces of West and East Flanders. There, some 20% of references to non-neuter nouns employs the neuter pronoun *het* 'it', exemplifying semantic agreement, whereas no trace is found of a tendency to use masculine *hij* 'he' to refer to neuter count nouns, as would be expected given the situation in northern Dutch.

At this point, it is not yet clear what has motivated these developments in Dutch gender. The collapse of masculine and feminine gender in the adnominal domain can be seen as the result of deflection, since Dutch has lost most of its adnominal morphology, such as its case system. As for pro-nominal gender, Audring (2006:113) proposes that the resemanticisation process is boosted by the fact that the noun phrase is underspecified with respect to the gender. Hence, in Audring's opinion the decreased visibility of adnominal gender plays a crucial role for the developments in pronomin-al gender. Data from the *World Atlas of Language Structures* (Corbett 2005), however, suggest another possibility. Typologically, apart from semantically motivated systems, gender systems are found in which both semantic and formal assignment rules play a role. Hence arbitrary gender systems are rare, indicating that arbitrariness of a gender system is a likely motivation for change. This, in turn, opens up the possibility that the devel-

opments in Dutch gender must be attributed to the fact that there is no formal ratio underlying gender assignment, which may have to do with the
loss of noun inflection, a phenomenon predating the loss of adnominal
morphology with several centuries. Indeed unlike in, e.g., German (Köpcke
and Zubin 1996) and French (Tucker, Lambert and Rigault 1977) most
gender assignments in present-day Dutch appear not to be formally motivated, and the most productive rule from Middle Dutch (i.e. nouns with
final -*e* have feminine gender; cf. Nijen Twilhaar 1992) has become obsolete since Dutch has dropped the final -*e* in most words. There is one observation corroborating the hypothesis that recoverability of gender on the
basis of nouns' form inhibits or at least delays changes in the gender system: in Dutch, some categories of derivations, e.g. feminine nouns on -*heid*
and -*nis* (e.g. *gezondheid* 'health', *ergernis* 'irritation') are among the
nouns most strongly resisting innovations in the gender system (Haeseryn
et al. 2002, §3.3.3).

2.2. Acquiring gender: German, English and Dutch

A landmark study on the acquisition of gender systems is Mills' (1986:113)
comparative study on German and English, in which the relevance of both
formal and semantic gender assignment rules is investigated. Her main
conclusion is that the speed with which children acquire a certain aspect of
the gender system "depends not on the categorization of the rules as semantic or formal, but rather on the relative 'clarity' of the rules in question
within the gender system." A relevant example comes from German. A
prominent gender rule in German is that words with final -*e* are feminine.
This rule is mastered by and large by children at the age of three years
(Mills 1986:70). Other rules that are present in adults but which have a
limited lexical scope or to which more exceptions occur (e.g., words on -/ft/
and -/cht/ are feminine, or words on -/et/ are neuter), still appear absent in
eight-year old children (Mills 1986:80).

   Mills' (1986) findings can be exploited for theorizing about the nature
of gender systems. For instance, a system like the German one, in which
some formal rules are already acquired at the age of three, can impossibly
be characterized as a predominantly semantic gender system. For Dutch,
however, evidence on early acquired formal rules is lacking. According to
De Houwer (1987:64-66) Dutch-speaking children of three years make use
of semantic rules for pronominal reference, such as the 'natural gender

rule': male humans are referred to using masculine pronouns such as *hij* 'he', female humans are referred to with feminine *zij* or *ze* 'she'. Masculine *hij* 'he' is also used for referents of which the natural gender is unknown or unimportant (e.g., some animals). The precise rules for reference to inanimates are not clear: both masculine *hij* 'he' and neuter *het* 'it' are used. Deviations from grammatical gender abound. In its semantic underpinnings, the pronominal gender system in Dutch three-year old children corresponds roughly to pronominal gender in young English speaking children (as described by Mills 1986:97-98). In addition, English and German strongly differ in the age at which children reach adult-like proficiency in their usage of pronominal gender. In German, gender mistakes are already quite rare at the age of 7, whereas deviations from the adult system still occur frequently in English-speaking children of that age. Thus, Mills (1986:97-98) finds many examples of masculine *he* referring to animals and count nouns (e.g. *The car is in the garage. HE is out of the rain*).

Apart from De Houwer's (1987), most data on the acquisition of Dutch gender concern adnominal gender rather than pronouns. In the adnominal domain, no semantically motivated deviations of grammatical gender have been reported. Dutch-speaking children appear to have more problems in acquiring adnominal gender than, for instance, French or German-speaking children. Unlike in French and German, Dutch-speaking children show a tendency to overgeneralise the common article *de* 'the' at the expense of neuter *het* 'the' (Van der Velde 2003:124-129), a tendency disappearing by the age of six. The problematic (L1 and L2) acquisition of Dutch adnominal gender is also discussed in Cornips and Hulk (2006), Hulk and Cornips (2006), and Blom, Polišenská and Weerman (2006). These findings indicate that there is probably no obvious systematicity underlying Dutch gender assignment; rather they confirm that Dutch gender is to a large extent arbitrary.

2.3. Research questions and methodology

From the data in 2.1 and 2.2 a number of research questions emerge, to which this paper hopes to provide an answer: first, are there differences in the way gender is acquired in the different regional varieties of Dutch? Second, in what respect does the acquisition of gender in these regional varieties resemble the acquisition of gender in English and German? And

third, can the acquisition data be used to predict the fate of gender in Dutch?

In order to answer these questions, data will be discussed from Dutch-speaking children aged seven or eight, an age in which almost no mistakes are made in the predominantly formally motivated German gender system, but quite many in the semantically motivated English system. Data are gathered for 19 Dutch children and 86 Belgian children. The Dutch children are from the province of Overijssel, a region in which the traditional three-gender system is lost. These children thus acquire an adnominal system such as the Standard Dutch one in Table 1, and feminine pronouns are no longer used to refer to historically feminine nouns. In contrast, the Belgian children all come from one of the most conservative areas in the Dutch language area as regards grammatical gender, the province of East Flanders. In this area, the dialectal system in Table 1 is commonly used in dialectal and even substandard speech. Also, pronominal gender is believed to be largely in line with the traditional tryadic system.

The choice of children of seven and eight allows collecting data by means of a written questionnaire. This entails that the investigated variety is Standard Dutch and not dialect, since children learn to read and write standard languages, not dialect.[3] The most important consequence of this is that no systematic variation can be expected as regards adnominal gender: although most southern dialects spoken in Belgium have preserved the traditional three-gender system, all varieties of Standard Dutch use the two-gender system in the adnominal domain, merely distinguishing between common gender (or *de*-words) and neuter (*het*-words). Hence the focus of this paper lies on pronominal gender, where there is no pressure whatsoever in the south to take over northern innovations (be it the dyadic grammatical system or the semantic one described in (1) above), and where the traditional three-way distinction between masculine, feminine and neuter gender is maintained. Hence differences can be expected in the way gender is used by the Dutch and the Belgian informants.

The written questionnaire consisted of sixteen or twenty four sentence completion tasks, depending on the version of the questionnaire. By using several questionnaires, it was possible to gather information for a larger number of words (n=39; see the appendix for a list of investigated items), without having to confront the children with a very long questionnaire. The test items were divided in four semantic categories that are relevant for pronominal gender (cf. *supra*): humans, animates, countable nouns, and mass nouns. The present selection of nouns allows for a test of Audring's

(2006) hypothesis in children, and also for detecting variation between different varieties of Dutch, but the selection is too narrow to be able to provide a full overview of all the semantic factors that may be at work. For instance, the questionnaire does not contain abstract nouns and collectives (see De Paepe and De Vogelaer 2008 and De Vos 2009, respectively, for discussion of these noun types). One relevant non-semantic factor is frequency (De Paepe and De Vogelaer 2008): infrequent items tend to trigger semantic agreement more often. This factor is kept constant in the present research: all items on the list are words that are frequently used by children (i.e. they are on the 'unanimity list' of Schaerlakens, Kohnstamm and Lejaegere 1999), which indeed occur in the varieties under scrutiny.

The East Flemish data are part of a larger, still ongoing investigation, in which more factors are included than in the present article. As a consequence, more lexical items are included in the East Flemish than in the Overijssel data. A comparison of the overall East Flemish results with the results for the nouns for which data are available for Overijssel too, yielded no significant differences, and thus below the overall results are given, since these are based on a larger data sample and thus more robust. Even within the more elaborate East Flemish survey (and in other investigations as well, cf. De Paepe and De Vogelaer 2008 and De Vos 2009), nouns within a certain semantic category behave rather uniformly (cf. also below). Hence resemantisation does not appear to show important lexical effects (apart perhaps from words for which grammatical gender shows geographical variation, but such words were kept out of the questionnaires).

Apart from noun semantics, agreement patterns in languages with semantic agreement also tend to depend on contextual and discourse factors (see, e.g., Curzan 2003:118-131 on Middle English and Siemund 2008 on present-day, non-standard varieties of English). Thus the test sentences with which pronouns are elicited may influence the results. In order to minimize variation resulting from differences between the test sentences, these all had a similar form. An example is given in (2): the children were instructed to fill in the pronoun (*hij* 'he', *ze* 'she' or *het* 'it') that they would use to refer to the bold-faced noun in a previous sentence (in this case: *bed* 'bed', a traditionally neuter noun triggering the use of *het* 'it' in most children). All pronouns to be filled in were subject pronouns. The bold-faced noun was always used as a sentence-initial subject in the first sentence, and was always preceded by a definite determiner, viz. a definite article or a possessive pronoun, and hence highly topical. These conditions are known to trigger the use of personal pronouns in the following sentence

(Gundel, Hedberg and Zacharski 1993; see also Comrie 1997 on Dutch). In addition, the second sentence was constructed in such a way that only the bold-faced noun could logically be referred to, and reference to the entire first sentence was equally unlikely. While Dutch, in general, shows quite a prolific use of demonstratives in subject position, this procedure proved successful, since indeed no demonstratives are found in the informants' answers.

(2)   Example sentence from the questionnaire
      Mijn bed staat in mijn kamer. ___ houdt me warm 's nachts.
      My bed is in my room. ___ keeps me warm at night.

In some cases, the determiner preceding the noun revealed gender informa-tion. According to De Paepe and De Vogelaer (2008:8), the presence of a gender-marking determiner does not affect the results: their questionnaire contained nouns for which both a test sentence was included with and without gender marking elements, and the results are identical. At the end of the task, the children were asked whether the test was easy or difficult, and they almost unanimously considered it an easy test.

3.   Acquiring pronominal gender in the Netherlands

3.1. Abandoning the grammatical three-gender system

Figure 1 displays the results for the 19 informants from the Netherlands. Proportions are shown of the answers in line with the (traditional) gram-matical gender of the questionnaire items per semantic type of noun, and per traditional gender. Clearly, the extent to which grammatical gender is used depends on the category of the noun that is referred to. Thus all mas-culine and feminine nouns with a human referent trigger the use of *hij* 'he' and *ze* 'she', respectively, while there is not a single *ze*-answer for the tradi-tionally feminine count nouns and mass nouns.
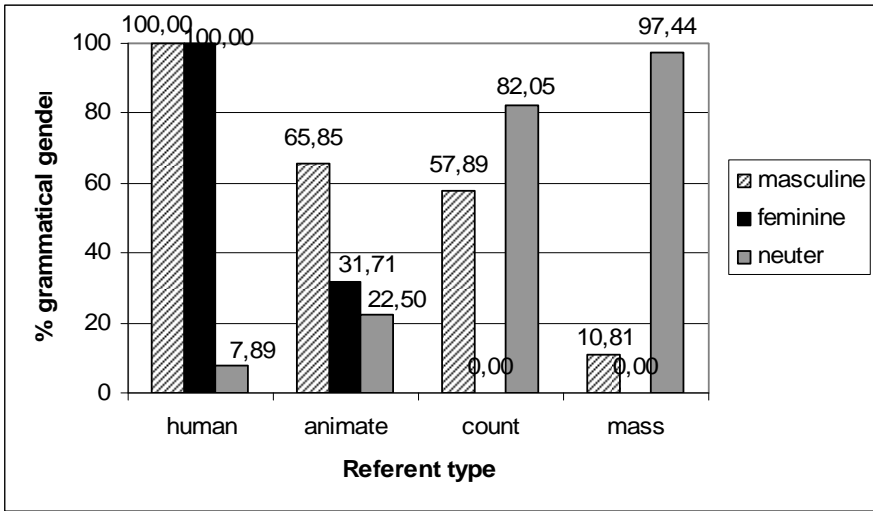
Figure 1. Grammatical gender in Dutch 7-8-year old children

The best way to assess to what extent the traditional three-gender system still applies, is a closer look into the number of attestations of the feminine pronoun *ze* 'she', since in the innovative systems, both the grammatical two-gender system and the semantic system described by Audring (2006), there is no motivation to use feminine pronouns other than reference to female referents. Apart from cases where there is indeed a semantic motivation for the use of *ze* 'she', such as in referring to female humans or, to a lesser extent, to female animals, the traditionally feminine nouns are never referred to with feminine *ze* 'she'. From this it can be concluded that the traditional three-gender system no longer plays a role.

Determining the extent to which the children's answers fit into a grammatical two-gender system or rather are semantically motivated is a more difficult task. On the one hand, the mere observation that the results for the different semantic categories differ strongly, indicates that noun semantics do play an important role. Strikingly, for most noun types for which grammatical gender appears to have been preserved quite well, grammatical gender corresponds to the alleged semantic gender in the innovative system described by Audring (2006) (see (1) above). In addition to the masculine and feminine nouns with human referents, these include the masculine animate nouns (65,85%), the masculine count nouns (57,89%), and the neuter mass nouns (97,44%). On the other hand, the fact that even these nouns

show variation (cf. the lack of 100% scores in the graph) evidences that the children's pronominal system cannot be captured in terms of a small number of categorical semantic rules. In addition, the high score for the neuter count nouns (82,05%) is clearly unexpected in light of Audring's thesis on (adult) northern Dutch, since this hypothesis predicts a higher number of attestations of *hij* 'he' than of *het* 'it' in this category. There are two ways to account for this result: either it reflects a semantically driven tendency to use *het* 'it' to refer to some count nouns not captured in Audring (2006), or it is the result of interference with grammatical gender, more precisely with the innovative, dyadic grammatical system (cf. *infra*).

3.2. Northern Dutch gender: a hybrid system

Table 2 not only shows the answers in which grammatical gender is preserved, it also provides an overview of all the other answers. Again the data are sorted according to the semantics of the noun that is referred to (human, animate, count nouns, mass nouns), and according to their traditional gender (masculine, feminine, neuter).

Table 2. Pronominal gender in 19 children from Overijssel (the Netherlands)

	MASCULINE			FEMININE			NEUTER		
	HIJ	ze	het	hij	ZE	het	hij	ze	HET
human:	38 100,00%	0 0,00%	0 0,00%	0 0,00%	38 100,00%	0 0,00%	12 31,58%	23 60,53%	3 7,89%
animate:	27 65,85%	7 17,07%	7 17,07%	25 60,98%	13 31,71%	3 7,32%	21 52,50%	10 25,00%	9 22,50%
count:	22 57,89%	2 5,26%	14 36,84%	19 47,50%	0 0,00%	21 52,50%	5 12,82%	2 5,13%	32 82,05%
mass:	4 10,81%	0 0,00%	33 89,19%	2 5,13%	0 0,00%	37 94,87%	1 2,56%	0 0,00%	38 97,44%

To a large extent these data confirm the picture emerging from the previous section: for three semantic categories (humans, animates and mass nouns), the use of gendered pronouns appears to be motivated semantically, i.e. in

line with natural gender on the one hand, and with a count/mass-distinction on the other. Thus, the vast majority of references to nouns denoting humans is with masculine *hij* 'he' or feminine *ze* 'she', even when the noun is grammatically neuter (e.g., *kind* 'child' or *meisje* 'girl'). The presence of *meisje* 'girl' in the questionnaire explains that the use of *ze* 'she' for neuter nouns denoting human referents exceeds 60%.

The high score of both *hij* 'he' and *ze* 'she' in the second row of Table 2 shows that natural gender also to a large extent drives reference to nouns denoting animates. Neither of these results is very surprising, as even Standard High German, a West Germanic variety in which the tryadic grammatical system has been maximally preserved, allows for these semantically motivated deviations from grammatical gender, albeit to a lesser extent (Mills 1986:51-53,93). Somewhat less expected is the non-marginal score of *het* 'it' for traditionally masculine (7/41 or 17,07%) and neuter animates (9/40 or 22,50%). The fourth row in the table shows the most straightforward results, in that virtually all references to mass nouns employ *het* 'it'. The scores range from 89,19% for traditionally masculine nouns to 97,44% for neuter nouns; the masculine pronouns are quite evenly distributed over the mass nouns in the questionnaire, viz. masculine *wijn* 'wine' (2 times *hij*) and *cola* 'coca-cola' (2), feminine *soep* 'soup' (1) and *melk* 'milk' (1), and neuter *zand* 'sand' (1). The data in the third row of the table are more chaotic: neither a dyadic grammatical system nor Audring's semantic system describes the situation very well.

Table 3. A grammatical two-gender system for count nouns?

	HIJ	HET
count, common:	41	35
count, neuter:	7	32

(*data extracted from Table 2 ; p<.001 (Chi square)*)

Table 3 orders the data from the third row assuming that the traditional three-gender system has been replaced by a two-gender grammatical system, in which common nouns (i.e. former masculine and feminine nouns) are referred to with *hij* 'he', and neuter nouns with *het* 'it'. Then a statistically significant correlation is found between grammatical gender and pronominal usage: common gender correlates with *hij* 'he', neuter gender with *het* 'it'. But it is immediately clear from Table 3 that the two-way grammatical gender distinction between common and neuter by no means categorically determines pronominal reference in the test data: deviations abound,

and they go in either direction, i.e. both the use of *het* 'it' for common nouns is found and the use of *hij* 'he' for neuter nouns.

But Table 3 also suggests that there is some asymmetry in these deviations: the attestations of *het* 'it' for common nouns by far outnumber the attestations of *hij* 'he' for neuter nouns. This use of *het* 'it' for masculine (14/38 or 36,84%) or feminine count nouns (21/40 or 52,50%) is especially remarkable since it is not captured in Audring's (2006) description of northern colloquial Dutch. Thus, while the first, second and fourth row might provide evidence in favor of a characterization of northern Dutch children's pronominal usage as being semantically motivated along the lines set out in Audring (2006), the results of the third row affect the overall applicability of her description. More specifically, there seem to be other factors than uncountability stimulating reference with *het* 'it'. These additional factors, however, cannot be easily defined. One possible factor may be abstractness: De Paepe and De Vogelaer (2008) find a tendency in Belgian children to refer with *het* 'it' to count nouns referring to abstract concepts, which may very well be present in Dutch children too. However, the questionnaire used for the present study did not contain any abstract items, but only count nouns with tangible referents, all of which trigger the use of *het* 'it' at least a couple of times: *mand* 'basket' (12 times *het*), *tafel* 'table' (9 times *het*), *stoel* 'chair' (9 times *het*), and *auto* 'car' (5 times *het*). Hence it seems as if in the language of the 7-8 year old Dutch children in the present study, *het* 'it' can in principle be used to refer to virtually every countable noun. The small proportions of *het* 'it' in the second row of Table 2 indicates that *het* 'it' can even be used for animates.

4.   Acquiring pronominal gender in Belgium

4.1. The persistence of the Belgian feminine

The Belgian data show quite a different picture. As for the preservation of grammatical gender, there are no noun classes where grammatical gender has been lost. For traditionally feminine count and mass nouns, where the innovative grammatical two-gender system uses *hij* 'he' and where there is no semantic motivation for the use of *ze* 'she' as well, slightly less than 40% attestations of *ze* 'she' are observed (36,84% and 37,21%, respective-

ly). For a number of other noun classes, too, grammatical gender scores more than 10% higher for the Belgian informants than in Figure 1, viz. for neuter nouns referring to humans (37,21%), for all nouns referring to animates (80,59%, 52,78%, and 41,28%), and for traditionally masculine count nouns (80,56%).

Overall, however, the results for the Belgian children show more similarities with the data from the Netherlands than one would expect on the basis of the differences observed in adults' language, where pronominal gender in the south is believed to be overwhelmingly in line with the traditional grammatical three-gender system. In Figure 2 it is indeed observed that different semantic categories clearly yield different results, testifying to the importance of noun semantics for pronominal gender. The categories for which grammatical gender is preserved well by and large correspond to Figure 1, and include classes in which grammatical gender does not conflict with natural gender, such as masculine and feminine nouns referring to humans and, to a lesser extent, animates. In addition, mass nouns seem to be strongly associated with the pronoun *het* 'it'. And here too, as in section 3.1, both masculine and neuter count nouns regularly trigger the use of the pronoun with according grammatical gender (with scores of 80,56% and 82,56%, respectively).
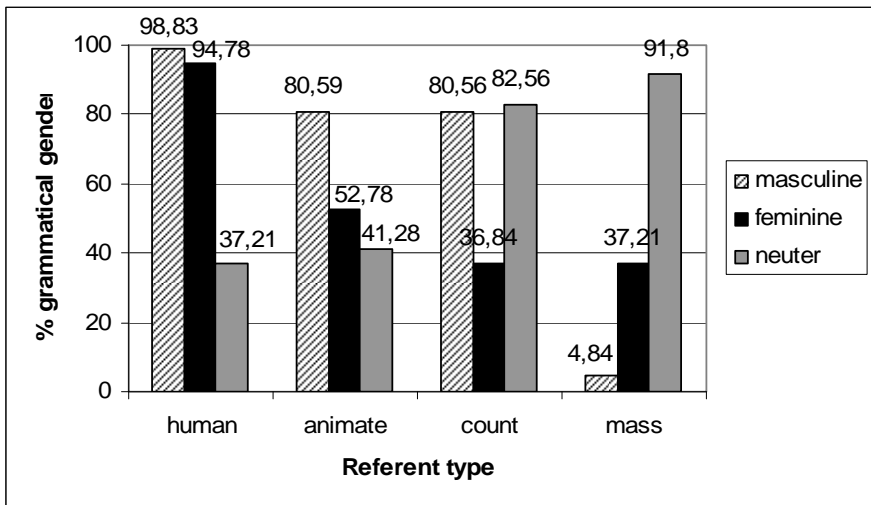
Figure 2. Grammatical gender in Belgian 7-8-year old children

4.2. Belgian Dutch gender and probabilistic grammars

Table 4, in which all the answers from the Belgian informants are given, confirms the picture emerging from Figure 2: the grammatical three-gender system does play a significant role, but many deviations are found. As in the Netherlands, reference to humans (row 1) and to a lesser extent to animates (row 2) appears to be in line with natural gender. Especially in animates, some nouns do not reveal information on natural gender, and in such cases children predominantly opt for *hij* 'he', although *ze* 'she' is used quite frequently as well. Mass nouns (row 4) present a clear preference for *het* 'it', and thus follow the system described by Audring (2006). As in the Overijssel data, the third row (count nouns) does not show the expected preference for *hij* 'he'.

Table 4. Pronominal gender in 86 children from East Flanders (Belgium)

	MASCULINE			FEMININE			NEUTER		
	HIJ	ze	het	hij	ZE	het	hij	ze	HET
human:	169 98,83%	1 0,58%	1 0,58%	6 4,48%	127 94,78%	1 0,75%	33 38,37%	21 24,42%	32 37,21%
animate:	137 80,59%	26 15,29%	7 4,12%	50 46,30%	57 52,78%	1 0,93%	46 42,20%	18 16,51%	45 41,28%
count:	145 80,56%	13 7,22%	22 12,22%	69 33,01%	77 36,84%	63 30,14%	8 9,30%	7 8,14%	71 82,56%
mass:	8 6,45%	11 8,87%	105 84,68%	13 15,12%	32 37,21%	41 47,67%	1 1,64%	4 6,56%	56 91,80%

Table 5 orders the data for the count nouns along the categories relevant in a dyadic grammatical system, i.e. a system distinguishing between common (i.e. traditionally masculine or feminine) and neuter gender. Indeed a correlation emerges between common gender and *hij* 'he', and between neuter and *het* 'it'. But here too, deviations are observed, as the pronoun *het* 'it' is frequently used to refer to non-neuter countable referents (12,22% for traditional masculines and 30,14% for feminines). Since this use of non-grammatical *het* 'it' clearly depends on the semantics of the antecedent (cf. increasing proportions of *het* as one moves down in Table 4), Table 4 and 5

show that Audring's (2006) description does not apply straightforwardly for pronominal reference by East Flemish children. Rather, the pronoun *het* 'it' seems to be available not only for lowly individuated mass nouns, but also for more strongly individuated referents, including count nouns and to some extent also animates (8 non-grammatical instances of *het* 'it', for masculine *stier* 'bull', *olifant* 'elephant' and *aap* 'monkey', and for feminine *koe* 'cow').

Table 5. A grammatical two-gender system for count nouns?

	HIJ	HET
count, common:	214	85
count, neuter:	8	71

(data extracted from Table 4 ; p<.001 (Chi square))

The overall conclusion is that in Standard Dutch as spoken by East Flemish children, three gender systems are operating: the traditional three-gender system, the innovative dyadic grammatical system, and semantic gender. Although there are differences with respect to the consistency with which each individual child applies grammatical or semantic gender, there are no children consistently using one system: all children show both answers which can only be explained as the result of grammatical gender, and answers pointing towards the use of semantic gender. In addition, there seems to be no regularity in the degree to which certain lexical items are liable to reference according to semantic or grammatical gender. Thus, in any given case, an East Flemish child may use semantic gender or grammatical gender. This has important consequences for the modeling of the gender system of East Flemish 7-8-year olds, which needs to be of a probabilistic nature rather than taking the form of rules determining when grammatical or semantic gender is used.

5.   Pronominal gender in German, English, and future Dutch

According to Mills (1986), the age at which an aspect of a language's gender system is acquired correlates with its relative clarity in the system. Hence, although semantic rules also play a role, German can be characterized as a language with grammatical gender: gender is primarily assigned on formal grounds, and three-year old children are already aware of the most important formal rule, i.e. the rule stipulating that words ending on a

schwa take feminine gender. Other rules are acquired as the child's lexicon expands (Mills 1986:85). In contrast, English speaking children only acquire semantic rules. A second difference between German and English concerns the speed with which children acquire the system: German-speaking children of seven and eight years make almost no gender mistakes, whereas English speaking children of the same age show a rather strong tendency to underuse *it* in referring to both animates and inanimates, which does not disappear before the age of ten (Mills 1986:91-92). Comparing the present data to Mills' (1986) data, the rather late acquisition of pronominal gender in both northern and southern Dutch provides a clear parallel to English. Also, although grammatical gender (either the two-way common-neuter distinction or the three way distinction between masculine, feminine and neuter) still has significant effects, the pronominal gender systems of both northern and southern 7-8-year-olds must be characterized primarily as semantic systems, as pronominal gender in English.

The fact that, unlike in German, the gender system of children speaking a southern variety of Dutch must be characterized as (largely) semantic in nature, indicates that, historically, noun semantics has gained importance in the south too. There are at least two reasons to believe that this tendency towards semantic gender will become even stronger in the future. The first indication is that the deviations from grammatical gender found in children persist in adolescence and adulthood, which makes it likely that the slow acquisition of grammatical gender will lead to language change in the long run (cf. Bybee and Slobin 1982). Relevant data are found in Geeraerts (1992) and especially in De Vos (2009), who also shows that many deviations from grammatical gender in adolescence and adulthood are indeed semantically motivated. The second indication is that, typologically speaking, simple, semantically motivated gender systems appear to be preferred over grammatical gender systems. More specifically, Nesset's (2006) 'Core Semantic Override Principle' predicts that both in northern and southern Dutch pronominal gender will develop in the same direction, viz. towards a semantic system.

More difficult to predict than the direction of future developments, is the precise timing. In general, the pace at which pronominal gender is resemanticised appears to be rather slow: at least in the dialects spoken in present-day West and East Flanders semantically motivated deviations from grammatical gender are quite rare in adults' language (De Vogelaer 2009). Hoppenbrouwers (1983), however, finds that in the Dutch province of Noord-Brabant there is a correlation between speakers' dialect proficiency and

their knowledge of the grammatical three-gender system. Hence processes of dialect loss and leveling may considerably accelerate the slow drift towards semantic gender, even though there is no direct pressure from Standard Dutch on the traditional pronominal gender system (cf. the fact that it is considered Standard Dutch, too). The explanation for this is probably that the more pervasive gender-marking morphology in the dialects offers dialect speakers more clues as to the grammatical gender of nouns, making it easier to acquire the system.

6.   Conclusions

Even when dealing with varieties of one language, viz. Dutch, the way in which the gender system is acquired mirrors the rule system that is used by adults (cf. Mills 1986 on German and English). In the present data, grammatical gender plays a much more important role in the southern, East Flemish child data than in the northern data from Overijssel. This is most clearly visible through the fact that East Flemish children show attestations of feminine gender for non-animate count nouns and mass nouns, whereas the children from Overijssel only use feminine pronouns to refer to female humans or animates. Essentially, however, both the northern and southern pronominal gender systems are acquired as predominantly semantic systems, thereby showing more similarities with English than with German. Hence it is likely that northern and southern Dutch pronominal gender will ultimately converge in a system of semantic agreement.

Notes

1.   Gunther De Vogelaer is a Postdoctoral Fellow of the Flemish Research Foundation (FWO) working at Ghent University. I would like to thank all people who have helped in gathering the data, viz. Leslie Buyle, Stefanie Ceelen, Isabelle De Clercq, Sietske Tilley and Lieve Troch for the Flemish data, and Sanneke Strous for the Dutch data. In addition, this paper has benefited from comments by the audience at the 'Cognitive Sociolinguistics'-workshop at the ICLC-conference in Krakow, July 2007.
2.   The feminine subject pronoun in Dutch is in principle *zij* rather than *ze*, which is the weak form. Strong pronouns, however, are not or seldom used to refer to nouns with non-human referents.

3. That said, there are no reasons to believe that investigating the children's most dialectal register would yield different results. Although some degree of hypercorrection has been reported (Geerts 1966:138-140), most speakers simply copy their intuitions on grammatical gender from their idiolect to their variety of Standard Dutch.

References

Audring, Jenny
   2006   Pronominal gender in spoken Dutch. *Journal of Germanic Linguistics* 18: 85-116.
Blom, Elma, Daniela Polišenská, and Fred Weerman
   2006   Effects of age in the acquisition of agreement inflection. *Morphology* 16: 313-336.
Bybee, Joan, and Dan Slobin
   1982   Why small children cannot change language on their own: suggestions from the English past tense. In *Papers from the 5th International Conference on Historical Linguistics*, A. Ahlqvist (ed.), 29-37. Amsterdam/Philadelphia: Benjamins.
Comrie, Bernard
   1997   Pragmatic binding: demonstratives as anaphors in Dutch. In *Proceedings of the Annual Meeting of the Berkeley Linguistic Society* 23: 50-61.
Corbett, Greville G.
   1991   *Gender*. Cambridge: Cambridge University Press.
   2005   Systems of gender assignment. In *The World Atlas of Language Structures*, Martin Haspelmath, Matthew Dryer, David Gil, and Bernard Comrie (eds.), 134-138. Oxford: Oxford University Press.
Cornips, Leonie, and Aafke Hulk
   2006   External and internal factors in bilingual and bidialectal language development: grammatical gender of the Dutch definite determiner. In *L2 Acquisition and Creole Genesis. Dialogues*, Claire Lefebvre, Lydia White, and Christine Jourdan (eds.), 355-378. Amsterdam/Philadelphia: Benjamins.
Curzan, Anne
   2003   *Gender Shifts in the History of English*. Cambridge: Cambridge University Press.
De Houwer, Annick
   1987   Nouns and their companions, or how a three-years-old handles the Dutch gender system. In *Belgian Journal of Linguistics 2: Perspec-*

tives on child language, Annick De Houwer, and Steven Gillis (eds.), 55-74. Amsterdam/Philadelphia: Benjamins.

De Paepe, Jessie, and Gunther De Vogelaer
    2008   Grammaticaal genus en pronominale verwijzing bij kinderen. Een taalverwervingsperspectief op een eeuwenoud grammaticaal probleem. *Neerlandistiek.nl* 08-02: 1-23.

De Schutter, Georges, Ton Goeman, Boudewijn van den Berg, and Thera de Jong
    2005   *Morfologische atlas van de Nederlandse dialecten. Deel 1: Meervoudsvorming bij zelfstandige naamwoorden, vorming van verkleinwoorden, geslacht bij zelfstandig naamwoord, bijvoeglijk naamwoord en bezittelijk voornaamwoord*. Amsterdam: Amsterdam University Press.

De Vogelaer, Gunther
    2009   Changing pronominal gender in Dutch: transmission or diffusion? In *Language variation, European perspectives II*, Stavroula Tsiplakou, Marilena Karyolemou, and Pavlos Pavlou (eds.), 71-80. Amsterdam/Philadelphia: Benjamins.

De Vos, Lien
    2009   Hersemantisering in het Nederlandse genussysteem. In *Papers of the Linguistic Society of Belgium* 3, Lobke Aelbrecht, Dany Jaspers, Frank Brisard, Patrick Dendale, and Bert Le Bruyn (eds.). Brussels: Linguistic Society of Belgium.
             http://webh01.ua.ac.be/linguist/SBKL/SBKL.htm

Geeraerts, Dirk
    1992   Pronominale masculiniseringsparameters in Vlaanderen. In *De binnenbouw van het Nederlands: een bundel artikelen voor Piet Paardekooper*, Hans Bennis, and Jan de Vries (eds.), 73-84. Dordrecht: ICG Publications.

Geerts, Guido
    1966   *Genus en geslacht in de Gouden Eeuw*. Brussels: Belgisch Interuniversitair Centrum voor Neerlandistiek.

Gundel, Jeanette, Nancy Hedberg, and Ron Zacharski
    1993   Cognitive status and the form of referring expressions in discourse. *Language* 69: 274–307.

Haeseryn, Walter, Kirsten Romijn, Guido Geerts, Jaap de Rooij, and Maarten van den Toorn
    2002   *Elektronische Algemene Nederlandse Spraakkunst*.
             http://www.let.ru.nl/ans/e-ans/

Hoppenbrouwers, Cor
    1983   Genus in een Brabants regiolect. *Tabu* 13: 1-25.

Hulk, Aafke, and Leonie Cornips
    2006   Neuter gender and interface vulnerability in child L2/2L1 Dutch. In *Paths of Development in L1 and L2 acquisition: In honor of Bonnie D. Schwartz*, Sharon Unsworth, Teresa Parodi, Antonella Sorace, Martha Young-Scholten, and Bonnie Schwartz (eds.), 107-134. Amsterdam/Philadelphia: Benjamins.

Köpcke, Klaus-Michael, and David Zubin
    1996   Prinzipien für die Genuszuweisung im Deutschen. In *Deutsch-typologisch. Institut für deutsche Sprache Jahrbuch 1995*, Ewald Lang, and Gisela Zifonun (eds.), 473-491. Berlin/New York: Walter de Gruyter.

Mills, Anne E.
    1986   *The Acquisition of Gender: A Study of English and German*. Berlin: Springer.

Nesset, Tore
    2006   Gender meets the usage-based model: four principles of rule interaction in gender assignment. *Lingua* 116: 1369–1393.

Nijen Twilhaar, Jan
    1992   Morfologische structuur en genusverandering I. *De Nieuwe Taalgids* 85: 302-318.

Plevoets, Koen, Dirk Speelman, and Dirk Geeraerts
    to appear De verspreiding van de -*e(n)*-uitgang in attributieve positie. *Taal en Tongval* theme issue 22 ('Perspectieven op het genus in het Nederlands').

Schaerlakens, Annemarie, Dolf Kohnstamm, and Marilyn Lejaegere
    1999   *Streeflijst woordenschat voor zesjarigen* [third, revised edtion]. Lisse: Swets and Zeitlinger.

Siemund, Peter
    2002   Mass versus count: pronominal gender in regional varieties of Germanic languages. *Sprachtypologie und Universalienforschung* (STUF) 55: 213-233.
    2008   *Pronominal Gender in English: A Study of English Varieties from a Cross-Linguistic Perspective*. London: Routledge.

Taeldeman, Johan
    1980   Inflectional aspects of adjectives in the dialects of Dutch-speaking Belgium. In *Studies in Dutch Phonology*, Wim Zonneveld, Frans Van Coetsem, and Orrin Robinson (eds.), 365-292. (Dutch Studies Vol. 4). Den Haag: Martinus Nijhoff.

Tucker, G. Richard, Wallace E. Lambert, and Andre Rigault
    1977   *The French Speaker's Skill with Grammatical Gender: An Example of Rule-Governed Behavior*. The Hague: Mouton.

Van der Velde, Marlies
  2003   Déterminants et pronoms en néerlandais et en français: syntaxe et acquisition. PhD. Diss, Université Paris 8.
Vousten, Rob
  1995   *Dialect als tweede taal: linguïstische enextra-linguïstische aspecten van de verwerving van een Noordlimburgs dialect door standaardtalige jongeren*. Amsterdam: Thesis publishers.

Appendix
Questionnaire items
(+ number of informants to whom the relevant item was presented)

The Netherlands
traditional masculines - human: jongen 'boy' (19), Harry Potter (19); animate: vlinder 'butterfly' (19), aap 'monkey' (19); count nouns: auto 'car' (19), stoel 'chair' (19); mass nouns: cola 'coca-cola' (19), wijn 'wine' (19)
traditional feminines - human: juf 'miss' (19), moeder 'mother' (19); animate: koe 'cow' (19), poes 'cat' (19); count nouns: tafel 'table' (19), mand 'basket' (19); mass nouns: soep 'soup' (19), melk 'milk' (19)
traditional neuters - human: kind 'child' (19), meisje 'girl' (19); animate: beest 'animal' (19), paard 'horse' (19); count nouns: bed 'bed' (19), boek 'book' (19); mass nouns: zand 'sand' (19), water 'water' (19)

Belgium
traditional masculines - human: jongen 'boy' (24), Samson (24), Harry Potter (62), Piet Piraat 'Pete the Pirate' (62); animate: vlinder 'butterfly' (24), stier 'bull' (24), olifant 'elephant' (62), aap 'monkey' (62); count nouns: auto 'car' (110), stoel 'chair' (24), fiets 'bike' (24), trommel 'drum' (24); mass nouns: cola 'coca-cola' (62), pijn 'ache' (62)
NB. the noun auto 'car' appeared twice on one of the versions of the questionnaire

traditional feminines - human: juf 'miss' (24), mama 'mother' (24), Roodkapje 'Little Red Riding Hood' (24), Assepoester 'Cinderella' (62); animate: koe 'cow' (24), kat 'cat' (86); count nouns: tafel 'table' (24), mand 'basket' (62), kast 'wardrobe' (62), zon 'sun' (62); mass nouns: soep 'soup' (24), melk 'milk' (62)
traditional neuters - human: kind 'child' (86); animate: beest 'animal' (24), lam 'lamb' (24), paard 'horse' (62); count nouns: bed 'bed' (24), feest 'party' (62); mass nouns: zand 'sand' (62)

Lectal variation in constructional semantics: "Benefactive" ditransitives in Dutch

Timothy Colleman

Abstract

This paper addresses the issue of lectal variation in constructional semantics through an exploration of the semantic constraints on the so-called "benefactive" ditransitive construction – i.e., the use of the ditransitive argument structure construction [Sbj [V Obj1 Obj2]] to encode events which involve a beneficiary rather than a (prototypical) recipient – in the two national varieties of Dutch. On the basis of corpus data, it will be shown that this construction occurs more frequently in Belgian than in Netherlandic Dutch. This difference in frequency will be accounted for in terms of a semantic "contiguity" constraint, which greatly reduces the lexical and semantic possibilities of the construction in (everyday registers of) Netherlandic Dutch.

Keywords: constructional semantics, ditransitive construction, benefactive ditransitive, regional variation, Construction Grammar, Belgian Dutch, Netherlandic Dutch, dative

1. Introduction*

One of the basic tenets of Construction Grammar and related constructionist approaches to language is that abstract grammatical constructions such as the ditransitive [Sbj [V Obj$_1$ Obj$_2$]] argument structure pattern are not fundamentally different from lexical items (or, in Construction Grammar terms, from *atomic lexically substantive* constructions). That is, just like lexical items, argument structure constructions are stored pairings of a certain form with a certain meaning. Hence, a lot of work in Construction Grammar is concerned with the elucidation of the semantics of abstract grammatical constructions. An often-quoted example is Goldberg's (1995) seminal analysis of the English ditransitive as a polysemous category built around a basic 'Agent successfully causes recipient to receive patient' sense.

It is well known that the semantic properties of lexical items are subject to various kinds of lectal variation, however. Yet, if one accepts that schematic argument structure constructions are meaningful linguistic entities in their own right just like lexical items, it is only to be expected that, on careful examination, such constructional meanings will be found to display a certain amount of lectal variation as well. This has not been a major focus of attention in Construction Grammar, but there are a number of studies which report on such intralingual differences in constructional semantics, including Wulff, Stefanowitsch and Gries (2007) on the distinct sets of verbs most typically associated with the *into*-causative in British versus American English and the associated constructional senses, Mukherjee and Hoffmann (2006) on the wider semantic range of the ditransitive in Indian English (as compared with standard British and American English) and Webelhuth and Dannenberg (2006) on the specific semantic properties of the "personal dative" construction in Southern American English.

The present study will address the issue of lectal variation in constructional semantics through an exploration of the semantic constraints on the so-called *"benefactive" ditransitive* construction in different (regional) varieties of Dutch. While the Dutch ditransitive construction has already attracted quite a lot of linguistic attention, the majority of existing studies are framed in the context of the well-known *dative alternation*. They focus on the relation between the ditransitive and so-called *prepositional dative* constructions in which the indirect object participant is marked by a preposition, usually *aan* for recipients and *voor* for beneficiaries, which are relevantly similar to the constructions with *to* and *for* in English, respectively (but see Colleman and De Clerck 2009). The main aim of such studies is to uncover the subtle semantic and/or discourse-pragmatic contrasts between these "competing" constructions (see e.g. Schermer-Vermeer 1991, Van Belle and Van Langendonck 1996, Janssen 1997, Colleman 2009a). In addition, the studies by Geeraerts (1998) and Colleman (2009b) examine the semantic range of the Dutch ditransitive in its own right from a constructionist perspective, providing overviews of the construction's various uses and subsenses. These studies do not deal with issues of lectal variation, however. In this respect, the present article is the first to look into the Dutch ditransitive from a combined constructionist and variational-linguistic perspective.

The paper is structured as follows. The following section introduces the "benefactive" ditransitive construction and summarizes existing claims about its semantic range in different regional varieties of Dutch. Section 3

presents the results of a small scale corpus investigation into the use of benefactive ditransitives in the two national varieties of the language, Netherlandic and Belgian Dutch, and Section 4 develops a hypothesis which accounts for the observed regional variation in terms of a semantic *contiguity* constraint. This discussion will involve a brief comparison with the equivalent English ditransitive construction. Section 5 briefly looks into another kind of lectal variation, viz. register variation, and Section 6 summarizes the main findings.

2.   On the "benefactive" ditransitive construction

2.1. Caused reception and benefaction

Many languages exhibit a three-participant argument structure construction which, in addition to all kinds of events in which an agent instigates a transfer of a patient towards a recipient, can also encode events which involve a *beneficiary* rather than a (prototypical) recipient as the third participant. Polish is a case in point: as shown by the examples from Dąbrowska (1997: 25–35) in (1) below, the Polish construction with dative and accusative objects can be used to encode prototypical transfer of possession events (1a–b) as well as events in which somebody carries out an action for the benefit of somebody else (1c–f). The dative object codes either the recipient or the beneficiary. For further discussion of the similarity between the encoding of events of caused reception and events of benefaction attested in many of the world's languages, see Shibatani (1996), Newman (1996: 95–97) and Kittilä (2005), *inter alia*.

(1) a. *Dał /Ofiarował*   *jej*   *obraz.*
         he gave/he presented   her:DAT   picture:ACC
         'He gave her a picture./He presented her with a picture.'

     b. *Wysłał / Podał*   *jej*   *książkę.*
         he sent/ he handed   her:DAT   book:ACC
         'He sent/handed her the book.'

     c. *Ala*   *uszyła* *mi*   *sukienkę.*
         Ala:NOM   sewed   me:DAT   dress:ACC
         'Ala sewed me a dress.'

d. *Magda   kupiła   Wojtkowi   książkę.*
  Magda:NOM   bought   Wojtek:DAT   book:ACC
  'Magda bought Wojtek a book.'

e. *Krystyna   otworzyła   Oli   drzwi*
  Krystyna:NOM   opened   Ola:DAT   door:ACC
  'Krystyna opened the door for Ola.'

f. *Krystyna   odrobiła   Oli   lekcje.*
  Krystyna:NOM   did   Ola:DAT   homework:ACC
  'Krystyna did Ola's homework for her'

As shown by some of the English glosses of the Polish examples, the English ditransitive or double object argument structure construction can also encode certain events of benefaction, next to events of caused reception. This use of the ditransitive argument structure pattern to encode events which involve a beneficiary rather than a (prototypical) recipient will henceforth be referred to as the *benefactive ditransitive* construction. It should be noted, however, that, semantically, this English benefactive construction is more restricted than the above mentioned Polish construction, as illustrated by the marked difference in acceptability between the clauses in (2a–b) and (2c–d) below.

(2)   a. Ala sewed me a dress.
      b. Magda bought Wojtek a book.
      c. *Krystyna opened Ola the door.
      d. *Krystyna did Ola the homework.

This can be described in terms of an "intended reception" constraint. Several authors have pointed out that for the ditransitive to be possible in English, the beneficiary has to be involved as the projected recipient of the patient. In (2a), for instance, Ala's action is aimed at the transfer of a sewed dress to the speaker (see, e.g., Allerton 1978, Wierzbicka 1986, Jackendoff 1990: 195–196, Langacker 1991: 360, Nisbet 2005). In terms of Kittilä's (2005) distinction between several subtypes of benefactive events, the English ditransitive only accommodates events of "recipient-benefaction", i.e. events in which the nature of benefaction is such that the beneficiary ultimately receives something by instigation of the agent. (2c–d) involve another kind of benefaction, viz. "deputive" or "substitutive" benefaction (i.e., the agent carries out an action *instead of* the beneficiary) and such substitutive beneficiaries cannot appear as the first object of the ditransitive

construction in English. Goldberg (1995) incorporates this view in her semantic network model of the English ditransitive: she distinguishes a sub-sense 'Agent intends to cause recipient to receive patient', which covers ditransitives with verbs of creation (including preparation) such as *make*, *bake*, *sew*, *cook*, etc. and with verbs of obtaining such as *buy*, *get*, *find*, etc.[1] The next subsection will show that in (standard Netherlandic) Dutch, the semantic possibilities of the benefactive ditransitive are (even) more limited than in English.

Before we proceed to a discussion of the Dutch construction and its lectal variation, however, it should be pointed out that the English construction has been shown to be subject to a certain degree of intralingual variation as well. Allerton (1978) presents the findings from a preliminary questionnaire study in which he asked fifty British university students to judge a series of forty test sentences of the pattern "*Could you V me NP?*". While examples such as *Could you pour me a cup of coffee?* and *Could you paint me a picture*? were accepted (virtually) unanimously and examples such as *Could you taste me this wine?* and *Could you teach me a class?* were quite consistently rejected, there were also a lot of cases which received mixed judgments. Some of these are listed in (3), with the assessment scores from Allerton (1978: 25) indicated between brackets.[2] The web examples in (4), quoted in Fellbaum (2005:223), illustrate that such clauses indeed occur in actual language use.

(3) a. Could you iron me these shirts? [76%]
     b. Could you wash me the dishes? [54%]
     c. Could you clean me the windows? [47%]
     d. Could you open me the door? [25%]

(4) a. Well, the rest is his story? Honey, can you iron me a shirt?
     b. You're a good boy, Joe. Now get busy and wash me some dishes.
         (Web examples quoted in Fellbaum 2005: 223)

Though the design of Allerton's study results in assessment scores which very probably hugely overestimate the acceptability of such clauses to the average native speaker of British English, his results clearly indicate that there is no clear-cut separation between possible and impossible benefactive ditransitives. This can be related to the fact that the above "intended reception" constraint comes with a certain amount of inherent fuzziness, i.e. whether a given event can be construed as involving intended causation of reception is a matter of degree rather than kind. In (3a), for instance, the

shirts probably already belonged to the beneficiary before the event, so there is no intended transfer of possession in the strict sense of the word. On the other hand, however, something previously unavailable to the benefici-ary *would* be made available to them as a result of the desired act, viz. some ironed shirts, and it would seem that this suffices for some speakers but not for others. It is perfectly possible that this variation within English (partly) follows regional lines, but, to my knowledge, this has not been systemati-cally investigated. We will briefly return to this observed lectal variation in English below.

2.2. Benefactive ditransitives in Dutch

The Dutch ditransitive construction is exemplified in (5) below. Just like its English counterpart, it combines a verb with a subject and two unmarked NP objects which, in the prototypical case, code the recipient and patient participants of a transfer of possession.

(5)   *Hij heeft de jongen een boek gegeven/overhandigd/verkocht.*
       He has the boy a book given/handed/sold.
       'He has given/handed/sold the boy a book.'

Many of the subsenses distinguished for the English construction in Gold-berg (1995) have their parallels in Dutch. In addition to canonical acts of giving, the construction can also be used to encode acts of prevention of possession (with verbs such as *weigeren* 'refuse' and *ontzeggen* 'deny'), various kinds of projected possessional transfers (e.g. with verbs such as *beloven* 'promise' or *aanbieden* 'offer'), communicative transfers (with verbs such as *vertellen* 'tell' and *meedelen* 'communicate'), etc. We refer to the above-mentioned studies by Geeraerts (1998) and Colleman (2009b) for overviews of the Dutch construction's semantic range.

One area in which this semantic parallelism between the English and Dutch ditransitive constructions does not hold, however, is exactly in the (im)possibility of encoding events of benefaction. The accepted view is that in present-day standard (Netherlandic) Dutch, the benefactive ditransitive is a very marked construction which is possible with a handful of rather infre-quent verbs related to food provision or preparation only, such as *(een drankje) inschenken* 'to pour (a drink)', *(een bord) opscheppen* 'to dish up

(a plate)' and in formal language also *(een maaltijd) bereiden* 'prepare (a meal)', see (6) for construed examples from the literature.

(6)   a. De ober heeft Piet een kop koffie ingeschonken
        (Schermer-Vermeer 1991: 219)
        'The waiter has poured Pete a cup of coffee.'

      b. Zal ik jou eens een lekker bord boerenkool opscheppen?
        (Schermer-Vermeer 1991: 219)
        'Shall I dish you up a tasty plate of borecole?'

      c. Ze had ons een heerlijke maaltijd bereid.
        (Haeseryn et al. 1997: 1165)
        'She had prepared us a delicious meal.'

In view of this, Verhagen (2002) correctly observes that standard Dutch does not have a productive benefactive ditransitive construction and marks the example in (7) as ungrammatical.

(7)   *Jan maakte haar een boterham. (Verhagen 2002: 415)
      'John made her a sandwich.'

Some comments are in order, however. First, it should be observed that in *earlier* phases of the language, the benefactive ditransitive was not at all limited to a handful of verbs of food provision or preparation of the kind illustrated in (6) above, but could be combined productively with all kinds of verbs to encode various subtypes of benefactive events. (8) lists a small number of attested examples from various periods. More examples can be found fairly easily by browsing the citations in the *Middelnederlandsch Woordenboek* [Middle Dutch Dictionary] (Verwijs and Verdam 1885-1952) and the *Woordenboek der Nederlandsche Taal* [Dictionary of the Dutch language] (De Vries, Te Winkel et al. 1882-1998).[3]

(8)   a. Daer cochte Joseph sinen here den vijften scoof vander vrucht.
        (c. 1300, Maerlant, Rhimed Bible)
        'There Joseph bought his lord the fifth part of the harvest.'

      b. Wat ic huer doe kin canse niet ghepaeyen.
        (16th C, anonymous ballade)
        'Whatever I do for her, I cannot content her.'

c. … dat hy hem de deur opende, en in een kamer voerde.
(N. Heinsius, De vermakelyken avanturier, 1695)
'… that he opened him the door, and led him to a room.'

d. [Ik] zal de deur aanwijzen, waarvoor zij mij de sleutel maken moet
(J.F. Oltmans, De Schaapherder, 1838)
'I'll point to the door for which she has to make me a key.'

On the basis of data from a corpus of literary fiction, Colleman (2002) concludes that in 19th Century Dutch, the construction could still be combined productively with verbs of creation and obtainment to encode events of recipient-benefaction.

Second, even in present-day Dutch, the semantic possibilities of the benefactive ditransitive are reported to be wider in a number of regional varieties of the language. According to Haeseryn et al. (1997: 1165), (construed) examples such as (9a) and (9b) below are typical of *Southern Dutch*, i.e. of language varieties spoken in Belgium and in the southern provinces of The Netherlands.

(9)   a. De hoogleraar kocht zijn vrouw een gouden armband.
          'The professor bought his wife a gold bracelet.'

       b.   Mijn vrouw heeft me een trui gebreid.
          'My wife knitted me a sweater.'

       (examples labeled as "regionally marked" in Haeseryn et al. 1997: 1165)

There have been a series of questionnaire and/or survey studies which corroborate that such uses indeed occur in *local dialect* and/or *regional substandard* varieties of southern, and as it happens also eastern, Dutch (see Van Bree 1981, Cornips 1994 and Colleman and De Vogelaer 2003). The present study looks into the possibilities of the benefactive ditransitive in the two *national* varieties of the language, i.e. Netherlandic vs. Belgian Dutch. Dutch is the standard language of about 16 million speakers in the Netherlands and about 6 million speakers in the northern part of Belgium (the region of Flanders). Historically, the Netherlandic standard was also adopted in Belgium. However, since the two speaker communities have been part of separate political entities for the largest part of the last four to five centuries, the Belgian variety of standard Dutch (which, in informal parlance, is often simply referred to as *Flemish*) is characterized by a number of linguistic differences from the Netherlandic standard. Often, these contrasts relate to characteristics of the southern Dutch dialect varieties

which are still widely used in everyday speech by many Belgian speakers. In the next section, we will investigate whether the wider semantic possibilities of the benefactive ditransitive in southern dialects manifest themselves in the standard language of Belgian speakers as well.

3. Benefactive ditransitives in Belgian vs. Netherlandic Dutch: A preliminary corpus investigation

In order to test the distribution of benefactive ditransitives in Netherlandic and Belgian Dutch, we selected six frequent verbs of either creation or obtainment, viz. *bouwen* 'build', *bakken* 'bake' and *tekenen* 'draw' (creation), and *kopen* 'buy', *halen* 'get, fetch' and *bestellen* 'order' (obtainment). None of these belong to the small set of verbs of the *inschenken* 'pour' type which can still be used ditransitively in the standard language according to the norm found in recent grammars such as Haeseryn et al. (1997). However, it is exactly with verbs from these two semantic classes, viz. creation and obtainment, that the benefactive ditransitive can be productively combined in many southern and eastern local dialect varieties, as shown in the above-mentioned studies.

Three different corpora were used, representing various modes and registers of standard Netherlandic and Belgian Dutch:

– the newspaper component of the CONDIV corpus of written Dutch (Grondelaers et al. 2000), with texts from three Dutch newspapers adding up to 4.8 million words and from four Belgian newspapers adding up to 12.7 million words;
– the Usenet component of the CONDIV corpus of written Dutch, with 7.7 million words of texts from Dutch discussion boards on the Internet (about various topics, including politics, sports, culture, etc., see Grondelaers et al. 2000 for an overview) and 5 million words from similar Belgian boards;
– the Corpus of spoken Dutch, with 5.7 words of spoken Netherlandic Dutch and 3.3 million words of spoken Belgian Dutch, representing various genres (spontaneous conversation, business meetings, radio shows, etc.; see, e.g., Van Eerten 2007).

As there is no way to automatically extract all clauses with ditransitive syntax from these corpora, I extracted all occurrences where a form of one of the six test verbs is combined with one out of a set of 22 personal, reflexive

and reciprocal pronouns, within a span of five words to the left or to the right (using the in-built search tool of the Corpus of Spoken Dutch and, for the CONDIV-corpus, the WordSmith Tools concordancer, Scott 2004).[5] The overwhelming majority of hits thus obtained were clauses in which the pronoun functions as the direct object, or in which the pronoun does not fill an argument position of the relevant verb at all, so the results had to be manually filtered to identify all instances in which the pronoun fills the indirect object position of the ditransitive construction. Needless to say, this procedure does not guarantee the retrieval of *all* benefactive ditransitives with the selected verbs from the corpora. First, the indirect object can take the form of a lexical NP rather than a pronoun, as illustrated in the Web example in (10) with the object *mijn pa* 'my dad'. However, the above-mentioned studies of the benefactive ditransitive in local dialect varieties of Dutch have shown that the construction is first and foremost combined with a pronominal indirect object.

(10)   'k was dit weekend om inkt geweest voor de epson die ik mijn pa voor
        zijn verjaardag gekocht had …
        <forum.belgiumdigital.com/archive/index.php/t-47056.html>
        'This weekend I went to buy ink for that Epson printer I bought my dad
        for his birthday.'

Second, the distance between the verb and the beneficiary object pronoun can be larger than five words. For the aim of the present study, however, a maximal distance of five words to the left or right was deemed sufficient, as this should suffice to give a good indication of the distribution of the construction over the various subcorpora.

    The results from the three corpora are merged in Table 1, which shows that the numbers of observed benefactive ditransitives are small. (For each verb, the table lists the total number of occurrences and the number of benefactive ditransitives.) *Kopen* 'buy' is the only verb with a fair number of ditransitive examples, mainly from the Belgian subcorpora. In fact, in all three investigated corpora, the proportion of ditransitive to other occurrences of *kopen* 'buy' is significantly larger in the Belgian subcorpus than in the Netherlandic subcorpus; see Table 2 for the detailed results.

Table 1. Observed benefactive ditransitives[6]

| | Netherlandic subcorpora | | Belgian subcorpora | |
	total	ditransitives	total	ditransitives
kopen 'buy'	11046	9	5914	75
halen 'get, fetch'	7131	1	7905	7
bestellen 'order'	992	0	954	3
bouwen 'build'	1803	0	2339	3
bakken 'bake'	469	(1)	554	1
tekenen 'draw'	965	0	1207	0

Table 2. Detailed results for *kopen* 'buy'

	ditransitive occurrences	other occurrences
CONDIV newspaper		
– Netherlandic subcorpus	3	860
– Belgian subcorpus	22	1986
		X^2 (N=2871, df=1) = 3.91; p = .0479
CONDIV Usenet		
– Netherlandic subcorpus	4	8236
– Belgian subcorpus	28	2762
		X^2 (N=11030, df=1) = 65.72; p < .0001
Corpus of spoken Dutch		
– Netherlandic subcorpus	2	1941
– Belgian subcorpus	25	1091
		X^2 (N=3059, df=1) = 37.01; p < .0001

For the other verbs, the frequencies are too small to allow meaningful statistical analysis, but it is noteworthy that virtually all observed ditransitive uses occur in the Belgian subcorpora. There are two exceptions: ditransitive *bakken* 'bake' is attested once in the Netherlandic part of the CONDIV newspaper corpus, and ditransitive *halen* 'get, fetch' is attested once in the Netherlandic part of the Corpus of spoken Dutch. Both of these "exceptional" Netherlandic examples merit some further discussion. First, compare the two ditransitive examples with *bakken* 'bake' in (11).

(11) a. Ik heb haar een doos noten gebracht en zij gaat mij nu een koek
      bakken.<forum.tuinadvies.be/8_2641_Walnoten_of_okkernoten.htm>
      'I took her a box of nuts and now she is going to bake me a cake'

b. Mogelijk hebben de bouwvakkers hem een poets willen bakken.
[CONDIV_newspaper_NL]
'Possibly, the construction workers wanted to play a joke on him.'

The Web example in (11a) is a genuine benefactive ditransitive: it describes an act in which the agent bakes something which is meant for the indirect object referent's consumption (and, presumably, benefit). The example from the Netherlandic part of the Corpus of spoken Dutch in (11b), by contrast, does not refer to an actual act of baking the product of which is intended for someone's benefit, but involves the fixed ditransitive expression *iemand een poets bakken* 'to play a trick on somebody' (the noun *poets* does not occur outside of this expression in present-day Dutch). Obviously, such idiomatic examples do not represent a productive use of the benefactive ditransitive construction but belong in a category of their own (which is why this *bakken* case is in parentheses in the Netherlandic column of Table 1).

The single attested Netherlandic example of ditransitive *halen* 'get, fetch' (12) occurs in an informal face-to-face conversation between two speakers from the province of Gelderland, in the east of the Netherlands, i.e. from one of the regions in which the ditransitive has been shown to be more flexible in the local dialects. Throughout this conversation (session code fn000969), the language of both speakers is characterized by various regionally marked features, the occurrence of ditransitive *halen* being just one of these.

(12) d'r zit ook nog een vaak een van die groene van die groene blaadjes moet
ik me d'ruit halen. [Corpus of Spoken Dutch_NL]
'Also, there is … often there is… those green… those tiny green leaves
which I have to remove from it (lit. which I have to get *me* out of it).'

In addition, the example in question is not a prototypical case of the benefactive ditransitive semantically, as the indirect object arguably does not code a real beneficiary: the speaker describes the process of preparing bean sprouts for consumption, which often involves removing green leaves from the product, but this cannot straightforwardly be construed as an act of (self)-benefaction. Probably, the example is better analysed as representing another non-standard use of the ditransitive pattern attested in some eastern dialects, in which an indirect object which is obligatorily coreferential with the subject serves to underscore the subject's agentivity (see Cornips 1994: 189–190 for related examples observed in Limburgian Dutch, and also see

Webelhuth and Dannenberg 2006 for an analysis of the Southern American construction in *He drank him a couple of beers* in similar "agentivity" terms). Compare with the semantically unproblematic example in (13), a case of reported speech attested in a spontaneous telephone conversation between Belgian speakers.

(13) *Schat wil 'k ik u nog een jeneverke halen?* [Corpus of spoken Dutch_B]
     'Darling, shall I get you another geneva?'

All of this suggests that Dutch verbs of creation and obtainment can be sub-divided in three groups according to their compatibility with the benefactive ditransitive construction:

– a small set of verbs related to food provision or preparation such as *in-schenken* 'pour', *opscheppen* 'dish up, ladle out' and *bereiden* 'prepare', which can be used with ditransitive syntax in both Belgian and Nether-landic Dutch;[7]
– the verb *kopen* 'buy', which is sporadically used ditransitively in Nether-landic Dutch but to a significantly larger extent in Belgian Dutch;
– verbs such as *halen* 'get, fetch', *bestellen* 'order', *bakken* 'bake', *bouwen* 'build', etc., which are sporadically used with ditransitive syntax in Bel-gian Dutch only.

The latter group constitutes an open class: in principle, any verb of creation or obtainment is eligible for occasional use in the benefactive ditransitive in Belgian Dutch. By way of illustration, (14) below lists a number of attested examples with verbs other than those listed in Table 1, from various Bel-gian sources.

(14) a. Zijn er nog vaklui te vinden die je een zwaard smeden van echte topkwaliteit? [De Morgen 20/08/2003]
        'Are there any craftsmen left who can forge you a top quality sword?'

     b. Die week liep ik door heg en steg in de gele trui die mijn grootmoeder me had gebreid.[Dimitri Verhulst, Dinsdagland]
        'That week I went walking in the countryside in the yellow sweater my grandmother knitted me.'

     c. Gorcha Davidova zocht zichzelf een leuk cadeautje voor haar verjaar-dag. < www.delaatsteshow.be/dlsblog2/wordpress/?p=212>

'Gorcha Davidova was looking for a nice present for her birthday.'
(lit. searched herself a nice present for her birthday)

d. In plaats van na te denken over hoe wij onze toekomstige families een
dak boven het hoofd gaan verdienen en intussen onze ouders én grootou-
ders hun senseopads en transcontinentale reizen zullen financieren. [De
Morgen 20/01/2007]
'Instead of thinking about how we are going to earn our future families a
roof over their heads, while at the same time our parents *and* grandpa-
rents will be financing their senseopads and transcontinental holidays.'

To sum up, with regard to the lexical possibilities of the benefactive ditran-
sitive construction, Belgian and Netherlandic Dutch are found to draw the
line at different places. The next section develops a hypothesis which ac-
counts for this observed regional variation in semantic terms.

4. A semantic "contiguity" constraint

4.1. Introduction: English versus Dutch

Section 2 above briefly discussed the intralingual variation observed in
English with regard to the acceptability of clauses like (15c) below, which
relates to the inherent fuzziness of the often-cited "intended reception"
constraint. While the clauses in (15a) and (15b) fulfil this constraint per-
fectly, the ironing example in (15c) presents a borderline case, hence the
mixed judgements.

(15) a. Ala sewed me a dress.
      b. Magda bought Wojtek a book.
      c. $^?$Could you iron me these shirts?

Obviously, this intended reception constraint cannot, by itself, account for
the intralingual variation observed in Dutch. Though each of the examples
in (16) below denotes a situation in which the indirect object referent is
unmistakably involved as the intended recipient of the direct object referent,
they are all of different status: (16a) represents a use which is generally
accepted in both Belgian and Netherlandic Dutch, while (16b) represents a
use which is regularly attested in Belgian Dutch but is marginal in Nether-

landic Dutch, and examples like (16c) are occasionally found in Belgian texts only.

(16) a. Albert wil jij me nog een glaasje wijn inschenken?
       [Corpus of spoken Dutch_NL]
       'Albert, can you pour me another glass of wine?'

      b. Volgens O. kocht de la Brassine hem ooit voor 100.000 frank nieuwe kleren in een "boetiek" aan de Anspachlaan.
       [CONDIV_newspaper_B]
       'According to O., de la Brassine once bought him 100,000 francs worth of new clothes in a "boutique" in the Anspach Avenue.'

      c. [I]k ga me een schuilkelder bouwen.
       [CONDIV_Usenet_B]
       'I'm going to build myself a bomb shelter.'

This goes to show that, compared to English, there must be *additional* semantic constraints at work in Dutch.

4.2. Two subevents

As a starting point for an elucidation of these additional constraints, consider Goldberg's (1995, 2002) semantic paraphrase of the ditransitive subsense instantiated by *I baked him a cake* etc. again: 'Agent intends to cause recipient to receive patient'. This paraphrase adequately fulfils Goldberg's aim to illustrate the semantic relatedness of benefactive ditransitives to other instantiations of the (English) ditransitive: the benefactive ditransitive is not a category *sui generis*, but fits nicely in the family of 'caused reception' senses displayed by the schematic ditransitive construction. On a more detailed level of analysis, however, a formulation in terms of intended causation of reception insufficiently distinguishes this benefactive subsense from other subsenses. In a way, for instance, clauses such as *Let's send him a letter* or *I'll throw you the ball* can be said to involve intended causation of reception as well (in that successful completion of the transfer is not guaranteed). The specific semantics of ditransitives with verbs of creation and obtainment is better captured by the following more elaborate paraphrase, adapted from Geeraerts (1998: 196): 'Agent carries out a preparatory action (involving patient), with a view to a subsequent transfer of patient to recipient'. This alternative paraphrase gives full justice to the two

subevents involved in this kind of benefaction event, viz. (i) the preparatory act, i.e. the creation or the obtainment, and (ii) the actual transfer of the product of this preparatory act to the recipient-beneficiary. Fillmore (2007: 134) makes a similar observation: in his terms, the ditransitive clause *I bought my sister a harmonica* "evokes a complex scenario involving two phases or sub-events: the purchase of the harmonica and its subsequent presentation to the speaker's sister".

4.3. The degree of contiguity required between the two subevents

4.3.1. *Restrictive varieties*

The hypothesis to be developed in this subsection is that the different regional varieties of Dutch form a cline with regard to the *degree of contiguity* required between the preparatory act and the actual transfer. In the most restrictive variety, i.e. standard Netherlandic Dutch, both subevents almost need to coincide for the ditransitive to be possible. This largely restricts the possibilities to cases such as *iemand een drankje inschenken* 'to pour sb a drink' and *iemand een bord groenten opscheppen* 'to dish sb up a plate of vegetables', i.e., to cases where the two subevents are virtually indistinguishable, so that the overall event could just as well be construed as a single act of giving. Schermer-Vermeer (1991: 219) hints at a similar explanation when she states that in (6a) and (6b) above, repeated here for convenience as (17), the verbs *inschenken* and *opscheppen* mean as much as "door middel van inschenken, respectievelijk opscheppen geven" [to give by means of pouring or dishing up].

(17) a. De ober heeft Piet een kop koffie ingeschonken.
       'The waiter has poured Pete a cup of coffee.'

      b. Zal ik jou eens een lekker bord boerenkool opscheppen?
       'Shall I dish you up a tasty plate of kale?'

There is independent evidence for the construal of such benefactive events of food provision as single acts of giving, in that *inschenken* 'pour' and *opscheppen* 'dish up, ladle out' can also be used in a number of *other* three-participant constructions which are normally reserved for verbs of giving.

First, consider the attested instances of *inschenken* and *opscheppen* used in the so-called *receptive* or *indirect-passive* construction with *krijgen* 'get' in (18) (cf. the structurally and semantically similar *kriegen/bekommen*-passive in German). Queries in the newspaper component of the Twente News Corpus (years 1994-2001) for any form of the verb *krijgen* combined with the past participle forms *ingeschonken* and *opgeschept* within a span of five words to the right or to the left, revealed nine receptive instances with *inschenken* and three receptive instances with *opscheppen*. Though these are of course small numbers – which is unsurprising in the light of the overall low frequency of the verbs in question – they suffice to show that the attested examples in (18a) and (18b) are not one-offs. Many more instances can be found on the Internet by Googling for strings such as *kreeg ingeschonken* 'got poured' or *krijgt opgeschept* 'gets dished up'.

(18) a. Wie melk uit een gekleurde fles kreeg ingeschonken, zou raar opkij-
       ken. [Reformatorisch Dagblad 15/01/2000]
       'Someone who was poured milk from a coloured bottle (lit. who got
       poured milk), would be much surprised.'

      b. Niet iedereen … zal erop staan dat hij zijn zabaglione ook daadwerke-
      lijk uit een peperdure massief koperen zabaglionepan krijgt opgeschept.
      [De Volkskrant 18/12/2001]
      'Not everybody will insist to be served his zabaglione (lit. to get dished
      up his zabaglione) from an expensive solid copper zabaglione pan.'

As shown in (19), this construction is compatible with verbs of giving such as *overhandigen* 'hand' or *bezorgen* 'deliver' as well as with closely related verbs such as *opsturen* 'send' or *aanbieden* 'offer', but *not* with ordinary verbs of creation or obtainment such as *kopen* 'buy', *halen* 'get, fetch', *bakken* 'bake', etc.

(19) a. De jongen kreeg een boek overhandigd/bezorgd/opgestuurd/ aangebo-
       den van zijn vader.
       'The boy was (lit. got) handed/delivered/sent/offered a book
       by his father.'

      b. *De jongen kreeg een boek gekocht/gehaald van zijn vader.
      'The boy was (lit. got) bought/fetched a book by his father.'

      c.   *De jongen kreeg een taart gebakken van zijn vader.
      'The boy was (lit. got) baked a cake by his father.'

Second, the Web examples in (20) illustrate the use of *inschenken* 'pour' in the prepositional dative construction with the default Dutch recipient preposition *aan* (cognate with English *on* but in this respect relevantly similar to English *to*).

(20) a. De barman … neemt het besluit om een 12,5 jaar oude whisky in te
          schenken aan de bargast.
          www.lachvandedag.com/1999/maart/26.htm
          'The bartender decides to pour a glass of 12,5 year old whisky "to" the
          customer'

       b. … en weer een ander was bereid de hele dag koffie en thee in te
          schenken aan de vele bezoekers.
          '… and another [volunteer] was willing to pour coffee and tea "to" the
          many visitors all day long'
          <www.passie.net/actueel.php?id=8>

Again, as shown in (21) below, this is a construction which accommodates all kinds of verbs of giving, but does not normally accommodate verbs of creation and obtainment (for further discussion of this *aan*-construction we can refer to Van Belle and Van Langendonck 1996 and Colleman 2009a, *inter alia*). If the recipient-beneficiary participant of verbs such as *kopen* 'buy', *maken* 'make', etc., is coded as a prepositional phrase, this is headed by *voor* 'for', not *aan* (see 21c).

(21) a. Jan geeft/verkoopt/overhandigt/stuurt een boek aan zijn vriend.
          'John gives/sells/hands/sends a book to his friend.'

       b. *Jan koopt/haalt/maakt een taart aan zijn vriend.
          'John buys/gets/makes a cake "to" his friend'

       c. Jan koopt/haalt/maakt een taart voor zijn vriend.
          'John buys/gets/makes a cake for his friend.'

Though the use of *aan* rather than *voor* in the examples in (20) above will very probably strike many speakers of Dutch as rather odd (the author of this article included), the relatively frequent occurrence of such examples on the WWW shows that, at least for some speakers, when it comes to preposition selection, *inschenken* 'pour' behaves like *geven* 'give' etc. rather than like *kopen* 'buy', *maken* 'make' and so on.[8]

To sum up, there are several indications that verbs such as *inschenken* 'pour' and *opscheppen* 'dish up, ladle out' are perceived by speakers as relevantly similar to verbs of giving rather than to ordinary verbs of creation/preparation or obtainment, and are treated accordingly. This semantic similarity to more central classes of ditransitive verbs explains why in the most restrictive varieties of Dutch these are the only verbs which can still be used with ditransitive syntax to encode an event of recipient-benefaction.

We conclude this subsection with an observation from Kirsner (1985: 251), cited approvingly in Schermer-Vermeer (1991: 220), *inter alia*, about the construed example in (22).

(22) ??Voordat Jan thuiskwam had Ineke hem al een borrel ingeschonken.
      'Before John came home, Irene had already poured him a drink.'

According to Kirsner, the ditransitive clause in (22) is awkward because John is not on the scene when the pouring occurs. This observation is in line with our hypothesis that for the ditransitive to be possible in standard Netherlandic Dutch, the preparatory act and the actual transfer must be contiguous, if not simultaneous, subevents.

4.3.2. More tolerant varieties

In the more tolerant varieties of Dutch, including supraregional Belgian Dutch, the ditransitive argument structure pattern can also be used to encode events of recipient-benefaction in which there is a looser bond between the two subevents. In contrast to the above instances with *inschenken* 'pour' and *opscheppen* 'dish up, ladle out', Belgian examples of the kind illustrated in (23) (= example 11a above) and (24) denote situations in which the preparatory action does not occur virtually simultaneously with the envisaged transfer.

(23) Ik heb haar een doos noten gebracht en zij gaat mij nu een koek bakken.
      <forum.tuinadvies.be/8_2641_Walnoten_of_okkernoten.htm>
      'I took her a box of nuts and now she is going to bake me a cake'

(24) Nu heb ik een verloopstekker gevonden bij m'n locale dealer, maar hij heeft em zelf nodig en hij is te lui om er mij enen te bestellen.
      [CONDIV_Usenet_B]

> 'I've found an adapter at my local dealer's now, but he needs it for himself and he is too lazy to order me one.'

Obviously, the act of baking of a cake in (23) must be completed *before* its product can be transferred to another person. Likewise, in (24), the action the speaker desires from his local hardware dealer, i.e. ordering an adapter, necessarily precedes the actual reception of this adapter by the speaker (possibly even by a couple of days or even weeks).

So, in Belgian Dutch, the ditransitive construction is more flexible with regard to the degree of contiguity required between the preparatory action and the transfer and as such can be used to encode a larger array of events of recipient-benefaction than the Netherlandic construction.[9] In fact, the Belgian Dutch ditransitive parallels the English construction in this regard: it is compatible with benefactive events as long as these involve an element of intended reception. It should be noted, however, that this is not to say that the benefactive ditransitives of Belgian Dutch and English are of equal status. In Stefanowitsch and Gries's (2003) collexeme analysis of the English ditransitive, two verbs of creation/preparation or obtainment, viz. *buy* and *cook*, turn out to belong to the 30 verbs most significantly attracted to the ditransitive construction, which suggests that the benefactive subsense associated with such verbs is quite well entrenched within the overall network of ditransitive subsenses. In Belgian Dutch, by contrast, the benefactive ditransitive constitutes a peripheral, infrequent subsense, as shown by the frequencies in Table 1: even *kopen* 'buy' is attested ditransitively in a mere 1,27% of its occurrences in the Belgian subcorpora (see Colleman 2009b for further discussion). In other words, the benefactive ditransitive covers a wider semantic range in Belgian Dutch than in Netherlandic Dutch, but even in the former variety it is infrequently attested in actual language use.

The observed in-between position of ditransitive *kopen* 'buy' – less usual in Netherlandic Dutch than, e.g., ditransitive *inschenken* 'pour', but not altogether impossible – is in accordance with the contiguity hypothesis as well. In 1903, the Dutch grammarian Den Hertog briefly discussed the semantic relation between (25a) and (25b) below, observing that the ditransitive clause in (25a) suggests that the daughter was present at the buying, while the *voor*-paraphrase in (25b) does not carry this suggestion (Den Hertog [1903] 1973: 46).

(25) a. Ik kocht mijn dochtertje een pop.
      'I bought my daughter a doll.'

b. Ik kocht een pop voor mijn dochtertje.
'I bought a doll for my daughter.'

This is in line with our hypothesis that for the ditransitive to be possible in standard Netherlandic Dutch, it matters whether the situation can be construed as a single act of giving: a buying event in which both the buyer and the beneficiary are present so that the beneficiary receives the object immediately upon the commercial transaction, is not unlike the food provision events discussed in the previous subsection in that there is a large degree of contiguity, if not overlap, between the preparatory action and the actual transfer. This motivates why the use of *kopen* 'buy' with ditransitive syntax, though rarely attested, is not altogether impossible in Netherlandic Dutch.

     To end this section, note that the ditransitive *kopen* 'buy' clause in (25a) denotes an event which is relevantly similar to prototypical acts of giving event in another key way as well in that it involves the *material* transfer of a concrete object. This seems to be another dimension of ditransitive semantics (besides contiguity) in which Belgian Dutch is more flexible than Netherlandic Dutch. Without delving deeper into this matter, it can be observed from the attested examples from the Belgian newspaper *De Morgen* in (26) that next to the fairly prototypical obtainment and creation events we have discussed so far, the ditransitive construction can also be used to encode events of recipient-benefaction which involve *abstract* rather than material reception in Belgian Dutch. The *roepen* 'call' event described in (26a) obviously does not involve a material transfer of Hitler to the coupists, nor does the *bedenken* 'think up, devise' clause in (26b) denote the material transfer of a remedy.

(26) a. Een commando van de putschisten trok eerst naar het rijkspropaganda-
          ministerie om Joseph Goebbels aan te houden, maar die zei doodgemoe-
          dereerd dat hij hen Hitler aan de telefoon kon roepen.
          [De Morgen 20/6/2004]
          'First, a squad of the coupists went to the Ministry of State Propaganda to
          arrest Joseph Goebbels, but he told them dead calm that he could call Hit-
          ler to the phone for them (lit. that he could call them Hitler to the phone).'

      b. Wanneer je me daar een remedie tegen kunt bedenken wil ik de vol-
          gende keer desnoods je snor trimmen. [De Morgen 28/10/2006]
          'If you can devise me a remedy for that, I am willing to trim your mous-
          tache next time, if need be.'

We leave it for future research to further test this additional semantic parameter and its exact relation to the contiguity parameter. This could be done by systematically comparing the assessments by Belgian and Netherlandic speakers of ditransitive *kopen* 'buy' clauses with concrete vs. abstract direct object referents and with absent vs. present indirect object referents, for instance. The next section presents another avenue for future research.

5.   Register variation

A verb which has not been discussed in the previous section is *bereiden* 'prepare'. Just like *inschenken* 'pour' and *opscheppen* 'dish up, ladle out', this verb is used with ditransitive syntax in Netherlandic Dutch as well as in Belgian Dutch, although examples such as (27) below violate the contiguity constraint just like examples with ditransitive *bakken* 'bake', *bestellen* 'order', *bouwen* 'build', etc. do. The preparation of a meal constitutes a subevent which is clearly separate from the actual "transfer" of this meal to its intended recipients.

(27)   Ze had ons een heerlijke maaltijd bereid. (Haeseryn et al. 1997: 1165)
        'She had prepared us a delicious meal.'

Two comments are in order, however. First, it should be noted that not all attested occurrences of ditransitive *bereiden* are genuine instances of the benefactive ditransitive on a par with the construed example in (27). In older phases of the language, *bereiden* was not only used as a verb of (food) preparation, but also as a formal alternative to *geven* 'give' or *verschaffen* 'provide' (see the lemma *bereiden* in the *Woordenboek der Nederlandsche Taal* [Dictionary of the Dutch language]). This older meaning has been preserved in a number of idiomatic uses, such as *iemand een warm(e) ontvangst/welkom bereiden* 'to give someone a warm welcome', *iemand een verrassing bereiden* 'to spring a surprise on someone' and *iemand genot/plezier bereiden* 'to give someone pleasure', see the examples in (28).

(28)   a. Groot-Brittannië wil af van zijn reputatie als een land dat   asiel-
            zoekers een warm welkom bereidt. [Het Parool 30/10/2001]
            'Great Britain wants to shed its reputation as a country which gives a
            warm welcome to asylum seekers.'

b. Het gepromoveerde Vada bereidde op de openingsdag van de compe-
titie veldkorfbal kampioen Blauw Wit een daverende verrassing (18-14).
[De Volkskrant 17/09/2001]
'On the first match day of the korfball competition, newly promoted Vada
sprang a big surprise on current champions Blauw Wit (18-14).'

c. Elke donderdagmiddag bezoekt hij de gracieuze, exotische Soraya, die
het weliswaar in dienst van een bureau voor geld doet, maar hem zoveel
plezier bereidt dat beiden van een uiterst bevredigende transactie kunnen
spreken. [De Volkskrant 10/09/1999]
'Every Thursday afternoon he visits the elegant, exotic Soraya who,
though she does it for money in the pay of an agency, gives him so much
pleasure that both can speak of a most satisfying transaction.'

Such idiomatic uses account for 19 of the 25 attested ditransitive examples
in the sample of 100 *bereiden* clauses from the TwNC (cf. footnote 7). The
remaining cases are genuine instances of the benefactive ditransitive, which
denote an act in which something (usually a meal) is prepared for some-
one's benefit, see (29) for two examples. Quite possibly the entrenchment
of the ditransitive idioms exemplified in (28) has been a factor in the pres-
ervation of the use instantiated by the clauses in (28) in Netherlandic Dutch.

(29) a. Zelfs nu de grandeur is vergaan en de spelers wachten op achterstallig
loon vloeit wijn en worden blikjes kaviaar opengeritst… Real bereidt zijn
Champions League-gasten   vanmiddag een prachtige lunch.
[Algemeen Dagblad, 22/11/1995]
'Even now, with the grandeur gone and players waiting for wage arrears,
the wine is flowing and tins of caviar are opened … *Real* prepares its
Champions League visitors a magnificent lunch.'

b. *Mevrouw Sinke noemt ook de actie Tafeltje-dek-je, waarbij bejaarden
die zichzelf geen warme maaltijd kunnen bereiden er dagelijks één thuis
krijgen bezorgd.* [Algemeen Dagblad, 10/05/1995]
'Mrs. Sinke also mentions the campaign *Tafeltje-dek-je*, in which elderly
people who cannot prepare themselves a hot meal, are delivered one at
home on a daily basis.'

Second, and most importantly, the verb *bereiden* in general and its ditransi-
tive use(s) in particular, belong to a fairly formal register in present-day
Dutch. Haeseryn et al. (1997: 1165) explicitly label the example in (27)
above as characteristic of formal Dutch (in contrast to similar examples
with e.g. *inschenken* 'pour').

This brings us to another dimension of lectal variation: the *regional* variation described in the previous sections is complemented by *register* variation. The contiguity constraint may go a long way towards explaining the distribution in everyday language, but, even in Netherlandic Dutch, the benefactive ditransitive is still used with some degree of productivity in certain text genres with a tendency for formal, archaic language. In view of the wider semantic possibilities of the construction in earlier phases of the language, this should of course not come as a surprise. By way of a brief illustration, consider the Web examples with *bouwen* 'build' in (30). As documented in Section 3 above, *bouwen* is not used with ditransitive syntax a single time in the Netherlandic subcorpora of CONDIV and the Corpus of spoken Dutch. Still, specialized Google queries in Dutch pages on the WWW return quite a lot of examples of the kind illustrated in (30), often from texts on religious, philosophical or esoteric subjects, or from poetry.[10]

(30) a. Hoeveel duidelijker is dan nu dit, dat wij Hem niet een tempel moeten
           bouwen, maar zelf een tempel moeten zijn.
           www.theologienet.nl
           'Then how much more obvious is this: that we should not build Him a
           temple, but rather be a temple ourselves.'

      b. Bouw me een cel met vier wanden licht
          Schenk me de vrijheid van je gezicht
          www.amnesty.nl/bibliotheek_vervolg/boek_eenbriefvanjou
          'Build me a cell with four walls of light
          Present me with the freedom of your face'

These examples instantiate the same grammatical construction as the everyday Belgian examples in (31), but, obviously, they represent an altogether different register of language.

(31) a. Ik wil mij een studioke bouwen en was zo eens aan het denken zijn er
          bepaalde online shops waar je studio matriaal kan kopen?
          <forum.belgiumdigital.com/showthread.php?t=24>
          'I'd like to build myself a little sound studio and was thinking like: are
          there online-shops where you can buy studio equipment?'

      b. Bouw me een pretpark, papa! [headline of an article about spoiled
          Hollywood kids in Het Nieuwsblad, 31/10/2005]
          'Build me an amusement park, daddy!'

To conclude, the examples in (30) suggest that in Netherlandic Dutch, benefactive ditransitives still occur more freely in certain formal text genres. It seems that the use of the ditransitive to encode non-contiguous events of recipient-benefaction has become a marker of conservative language in this national variety and that it is thus only tolerated in highly formal and/or archaic registers. The exact size of its presence there will have to be determined by research involving specialized corpora.

6.   Conclusions

The above has shown that present-day Dutch is subject to lectal variation in the use of the ditransitive argument structure construction to encode events which involve a beneficiary rather than a prototypical recipient as a third participant. In earlier phases of the language, the ditransitive could encode various subtypes of benefactive events. In everyday standard Netherlandic Dutch, however, this benefactive use of the ditransitive construction is heavily constrained. In addition to the "intended reception" constraint known from the relevant literature on English, the Netherlandic construction is subject to a "contiguity" constraint, to the effect that it occurs with a handful of verbs of food provision of the *inschenken* 'pour'-type only. In Belgian Dutch, which is generally taken to be the more conservative variety, the semantic possibilities are wider. Just like the corresponding English construction, the benefactive ditransitive of Belgian Dutch can be productively combined with verbs of creation/preparation and with verbs of obtainment to encode events of recipient-benefaction, though it has also been shown that such uses are quite infrequent. In addition, we have seen that the observed regional variation is complemented by register variation. Even in Netherlandic Dutch, the benefactive ditransitive is still used with a certain degree of productivity in text genres with a tendency for formal, archaic language. All of this suggests that we are dealing with a process of semantic retraction: the benefactive ditransitive constitutes a peripheral use of the ditransitive construction which is in the process of disappearing from the grammar, but at various speeds in the various varieties of the language.

On a more general level, the case of the Dutch benefactive ditransitive illustrates that the semantic properties of abstract argument structure constructions can be subject to language-internal variation just like the semantic properties of lexical items can. The incorporation of such patterns of lectal variation in models of constructional semantics may prove an impor-

tant contribution to the further development of Construction Grammar as an overall theory of language.

Notes

1.   Kay (2005) presents an alternative Construction Grammar analysis, analyzing the benefactive ditransitive as a construction in its own right rather than a subconstruction of the schematic ditransitive. The pros and cons of these two positions need not concern us here.

2.   The scores should not simply be read as "accepted by N%" of informants. Allerton asked his 50 informants to rate the sentences as "acceptable", "marginal" or "impossible". These assessments were weighted as 1, 0.5, and 0, respectively, and the overall scores were converted to percentages.

3.   Note that, technically, Middle Dutch examples such as (8a) do not involve the ditransitive construction (in its narrow sense of a construction with two zero-marked NP objects) but rather its predecessor in Dutch grammar, viz. the construction with accusative and dative objects (*sinen here*, for instance, is overtly marked for dative case). The possibility to encode benefactive events with this pattern, however, survived long after the case distinctions were lost.

4.   For a concise English introduction to the linguistic situation in Belgium and the Netherlands, see, e.g., De Schutter (1994).

5.   This group of 22 pronouns consisted of: (i) the full and reduced object forms of the personal pronouns of all grammatical persons, with the exception of the 3rd person neuter form *het* 'it', i.e. *mij, me, jou, je, u, hem, haar, ons, jullie, hen, hun, ze*; (ii) the weak and strong forms of the reflexive pronouns of all grammatical persons, i.e. *mezelf, mijzelf, jezelf, jouzelf, uzelf, zich, zichzelf, onszelf* and (iii) the reciprocal pronoun *elkaar* 'each other' and its informal variant *mekaar*.

6. The total frequencies mentioned for each of the six verbs are estimates based on the overall results of automatic searches for their infinitive and inflected forms.

7. The overall frequencies of these verbs are too low to test this claim by means of the same corpus method used for the verbs in Table 1. However, their compatibility with the benefactive ditransitive even in Netherlandic standard Dutch is uncontroversial, as shown by data from a larger corpus. Samples of 100 randomly selected *inschenken* 'pour' and *bereiden* 'prepare' occurrences from the *Twente News Corpus* (henceforth: TwNC, a large corpus of contemporary Dutch newspaper texts) contained 27 and 25 ditransitive instances, respectively. For *opscheppen* 'dish up, ladle out', a similar sample of 100 relevant occurrences is less easily obtained, as the verb is used in the synchronically unrelated sense 'brag, boast' in the overwhelming majority of its occurrences, but specialized queries in the TwNC return a number of examples such as (i) below, which corroborate that such uses occur in Netherlandic Dutch.
(i) Zijn moeder, een klein sherpa-vrouwtje met lange rokken, schept hem troostend nog wat dhal bat op. [De Volkskrant 06/05/2000]
'Comfortingly, his mother, a tiny sherpa woman in long skirts, ladles him out some more *dhal bat*.'

8. Google queries for any non-split form of the verb *inschenken* 'pour' *immediately followed* by the preposition *aan* on 01/07/2009 illustrate this relatively frequent occurrence with a respectable 23 contemporary examples of the construction illustrated in (20). I have not been able to find similar *aan*-examples with the verb *opscheppen* 'dish up, ladle out'. It is reasonable to assume, however, that speakers who find the *inschenken* examples in (20) acceptable, will also accept *aan*-clauses with *opscheppen*.

9. This statement abstracts away from possible regional differences *within* Belgian Dutch. Van Bree (1981) and Colleman and De Vogelaer (2003) have shown that in regional dialects from the Belgian province of West Flanders, for instance, the benefactive ditransitive is largely restricted to *inschenken* 'pour' etc., just like in standard Netherlandic Dutch. It is quite possible that in their supraregional communication as well, speakers from West Flanders will use fewer benefactive ditransitives, with fewer verbs, than speakers from, say, East Flanders or Belgian Limburg. Testing this hypothesis is a topic for future research.

10. Google queries (on 15/5/2007) within the domain .nl for all strings of any form of *bouwen* immediately followed by one of the personal pronouns *me* 'me', *hem* 'him' or *zich* 'oneself' immediately followed by the indefinite determiner *een* 'a' returned 41 ditransitive clauses of the kind illustrated in (30).

References

Allerton, David J.
    1978   Generating indirect objects in English. *Journal of Linguistics* 14: 21-33.
Colleman, Timothy
    2002   De benefactieve dubbelobject-constructie in het 19de-eeuws Nederlands [The benefactive double object construction in 19th Century Dutch]. In *De taal in Vlaanderen in de 19de eeuw. Historisch-sociolinguïstische onderzoekingen* [The language of Flanders in the 19th Century. Historical-sociolinguistic investigations], Roland Willemyns (ed.), 509-528. Theme volume of *Verslagen en mededelingen van de Koninklijke Academie voor Nederlandse Taal- en Letterkunde*.
    2009a   Verb disposition and argument structure alternations: a corpus study of the Dutch dative alternation. *Language Sciences* 31: 593-611.
    2009b   The semantic range of the Dutch double object construction. A collostructional perspective. *Constructions and Frames* 1(2): 190-221.
Colleman, Timothy, and Bernard De Clerck
    2009   'Caused Motion'? The semantics of the English *to*-dative and the Dutch *aan*-dative. *Cognitive Linguistics* 20: 5-42.
Colleman, Timothy, and Gunther De Vogelaer
    2003   De benefactiefconstructie in de zuidelijk-Nederlandse dialecten [The benefactive construction in southern Dutch dialects]. *Taal en Tongval* 15/16: 184-208.
Cornips, Leonie
    1994   *Syntactische variatie in het Algemeen Nederlands van Heerlen* [Syntactic variation in the standard Dutch of Heerlen]. Amsterdam: IFOTT.
Dąbrowska, Ewa
    1997   *Cognitive Semantics and the Polish Dative*. Berlin/New York: Mouton de Gruyter.
Den Hertog, Cornelis H.
    1973   *Nederlandse Spraakkunst. Eerste stuk: De leer van de enkelvoudige zin* [Dutch grammar. Part I: The study of the simple clause]. Edited by H. Hulshof. Amsterdam: Versluys. Original edition, Amsterdam: Versluys, 1903.
De Schutter, Georges
    1994   Dutch. In *The Germanic Languages*, Ekkehard König, and Johan van der Auwera (eds.), 439-477. London/New York: Routledge.

De Vries, Matthias, Lammert Te Winkel et al. (eds.)
  1882-1998 *Woordenboek der Nederlandsche Taal* [Dictionary of the Dutch Language]. 29 volumes. 's-Gravenhage/Leiden etc.: M. Nijhoff/A.W. Sijthoff etc.
Fellbaum, Christiane
  2005 Examining the constraints on the benefactive alternation by using the World Wide Web as a corpus. In *Evidence in Linguistics: Empirical, Theoretical, and Computational Perspectives*, Marga Reis and Stephan Kepser (eds.), 209-240. Berlin/New York: Mouton de Gruyter.
Fillmore, Charles
  2007 Valency issues in Framenet. In *Valency: Theoretical, Descriptive and Cognitive Issues*, Thomas Herbst, and Karin Götzz-Votteler (eds.), 129-160. Berlin/New York: Mouton de Gruyter.
Geeraerts, Dirk
  1998 The semantic structure of the indirect object in Dutch. In *The Dative. Volume 2: Theoretical and Contrastive Studies*. Willy van Langendonck, and William Van Belle (eds.), 185-210. Amsterdam/Philadelphia: Benjamins.
Goldberg, Adele E.
  1995 *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.
  2002 Surface generalizations: An alternative to alternations. *Cognitive Linguistics* 13: 327-356.
Grondelaers, Stefan, Katrien Deygers, Hilde Van Aken, Vicky Van den Heede, and
Dirk Speelman
  2000 Het CONDIV-corpus geschreven Nederlands [The CONDIV corpus of written Dutch]. *Nederlandse Taalkunde* 5: 356-363.
Haeseryn, W., K. Romijn, G. Geerts, J. de Rooij, and M.C. van den Toorn
  1997 *Algemene Nederlandse spraakkunst. Tweede, geheel herziene druk* [General Dutch Grammar. Second, completely revised edition]. Groningen: Martinus Nijhoff/Deurne: Wolters Plantyn.
Jackendoff, Ray S.
  1990 *Semantic Structures*. Cambridge, MA: The MIT Press.
Janssen, Theo
  1997 Giving in Dutch: an intra-lexematical and inter-lexematical description. In *The Linguistics of 'Giving'*, John Newman (ed.), 267-306. Amsterdam/Philadelphia: Benjamins.
Kay, Paul
  2005 Argument structure constructions and the argument-adjunct distinction. In *Grammatical Constructions: Back to the Roots*, Mirjam

Fried, and Hans Boas (eds.), 71-98. Amsterdam/Philadelphia: Benjamins.

Langacker, Ronald W.
1991   *The Foundations of Cognitive Grammar. Volume. 2: Descriptive Application*. Stanford: Stanford University Press.

Kirsner, Robert S.
1985   Iconicity and Grammatical meaning. In *Iconicity in Syntax*, John Haiman (ed.), 249-270. Amsterdam/Philadelphia: Benjamins.

Kittilä, Seppo
2005   Recipient-prominence vs. beneficiary-prominence. *Linguistic Typology* 9: 269-297.

Mukherjee, Joybrato, and Sebastian Hoffmann,
2006   Describing verb-complementational profiles of new Englishes: a pilot study of Indian English. *English World-Wide* 27: 147-173.

Newman, John
1996   *Give. A Cognitive Linguistic Study*. Berlin/New York: Mouton de Gruyter.

Nisbet, Tim
2005   Benefactives in English: evidence against argumenthood. *Reading Working Papers in Linguistics* 8: 51-67.

Schermer-Vermeer, Ina
1991   *Substantiële versus formele taalbeschrijving: het indirect object in het Nederlands* [Substantial versus formal language description. The indirect object in Dutch]. Amsterdam: Universiteit van Amsterdam.

Scott, Mike
2004   *WordSmith Tools version 4*. Oxford: Oxford University Press.

Shibatani, Masayoshi
1996   Applicatives and benefactives: a cognitive account. In *Grammatical Constructions: Their Form and Meaning*, Masayoshi Shibatani, and Sandra A. Thompson (eds.), 157-194. Oxford: Oxford University Press.

Stefanowitsch, Anatol, and Stefan Th. Gries
2003   Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8: 209-243.

Van Belle, William, and Willy Van Langendonck
1996   The indirect object in Dutch. In *The Dative. Volume 1: Descriptive studies*, William Van Belle, and Willy van Langendonck (eds.), 217-250. Amsterdam/Philadelphia: Benjamins.

Van Bree, Cornelis
1981   *Hebben-constructies en datiefconstructies binnen het Nederlandse taalgebied. Een taalgeografisch onderzoek* [*Hebben*-constructions and dative constructions in the Dutch language area. A geolinguistic investigation]. Leiden: Rijksuniversiteit Leiden.

Van Eerten, Laura
    2007   Over het Corpus Gesproken Nederlands [On the Corpus of Spoken Dutch]. *Nederlandse Taalkunde* 12: 194-215.

Verhagen, Arie
    2002   From parts to wholes and back again. *Cognitive Linguistics* 13: 403–439.

Verwijs, Eelco, and Jakob Verdam
  1885-1952 *Middelnederlandsch Woordenboek* [Middle Dutch Dictionary]. 11 volumes. 's-Gravenhage: Martinus Nijhoff.

Webelhuth, Gert, and Clare Dannenberg
    2006   Southern American English personal datives. The theoretical significance of dialectal variation. *American Speech* 81: 31-55.

Wierzbicka, Anna
    1986   The semantics of 'internal dative' in English. *Quaderni di Semantica* 7: 121-135.

Wulff, Stefanie, Anatol Stefanowitsch, and Stefan Gries
    2007   Brutal Brits and persuasive Americans: Variety-specific meaning construction in the *into*-causative. In *Aspects of Meaning Construction in Lexicon and Grammar*, Günter Radden, Klaus-Michael Köpcke, Thomas Berg, and Peter Siemund (eds.), 265-281. Amsterdam/Philadelphia: Benjamins.

Part three.
Variation of lectal awareness and attitudes

Lectal acquisition and linguistic stereotype formation: an empirical study

Gitte Kristiansen

Abstract

This chapter reports on an empirical study on child language acquisition conducted in 2007. The overall aims of the investigation were to a) address the question of *when* young children acquire (receptive and active) competence of lectal varieties and b) discuss potential predictors of the success rate: *how* do children acquire this knowledge? How do they learn to correlate tokens with types as effectively as they do? Adopting a usage-based approach to lectal stereotype formation we predict that structure is determined by use and that purely formal characteristics (such as phonetic salience and contrast) will have less effect on correct identification than relative social salience (such as social stereotyping).

The chapter evolves in five main steps. In the Introduction we briefly provide an outline of the theoretical background within which the experiment is embedded. Section 2 describes the design and the results of the first experiment conducted, which aimed to assess the degree of identifiability of L1 accents in young children. Section 3 in turn outlines the design and the findings of the second experiment, on correct identification of L2 accents by the same age group as in experiment 1. In section 4 we address the explanatory dimension by analyzing the data collected in a questionnaire distributed to the subjects. Finally, we discuss the theoretical implications of the findings.

Keywords: language acquisition, cognitive dialectology, linguistic stereotype formation, lectal disambiguation, template-based categorization

1. Introduction

In previous publications (Kristiansen 2001, 2003, 2006, 2008) I have argued in favor of a usage-based prototype-theoretical approach to lectal categorization, perception and awareness. Viewing lectal variation (e.g. variation at the level of dialects, accents, registers, styles etc.) from the

perspective of a usage-based cognitive dialectology I have been particularly interested in social cognition, the extent to which phonetic detail is perceived in folk perception and the implications for cognitive dialectology and phonology.

In general terms I have viewed lectal varieties in terms of prototype categories and argued (cf. Bybee 2001; Pierrehumbert 2001) that hearers store and make sense of linguistic detail at the level of phonetics in order to navigate effectively in the social dimension: intraphonemic and transphonemic salient contrasts serve to establish exophoric reference to social categorizations and allow for hearers to categorize speakers on the basis of their speech patterns. Accents are socially diagnostic because linguistic stereotypes (sets of fairly abstract linguistic schemata capturing the essence of what an out group speaks like, or more technically speaking: central images of lectal categories) emerge on experiential grounds in early childhood and from then on more and more effectively evoke the social categorizations they originated from. These serve as shortcuts to identification (i.e. from linguistic stereotype to social category: where is this speaker from?) and characterization (i.e. from social category to social stereotype: what is this speaker like?)

In more specific terms, I have argued that a perspective according to which hearers do have receptive competence of lectal varieties have both theoretical and practical implications for cognitive phonology. In several publications (Kristiansen 2003, 2006, 2008) I have called for a usage-based analysis that will serve to incorporate the social dimension into the theoretical framework and the descriptive analyses to a greater extent

In this chapter we report on a set of studies carried out in 2007 aiming at a) examining *when* young children acquire (receptive and active) competence of lectal varieties and b) discussing possible predictors of the success rate: *how* do children acquire this knowledge? How do they learn to correlate tokens with types as effectively as they do?

The interest in the first dimension, which is predominantly descriptive, stems from a debate in cognitive phonology about the existence of passive competence. As lectal categorization and perception are central to cognitive dialectology, which is the main topic of this chapter, I was puzzled a number of years ago to encounter divergent opinions on the part of scholars in cognitive phonology regarding the existence of lay knowledge about phonetic detail (cf. Kristiansen 2003, 2006; Nathan 1996, 2006). For someone with an interest in the meaningfulness of linguistic structure - and in Cognitive Linguistics as opposed to most other disciplines, meaningfulness is not

relegated to content words such as verbs or nouns, but pertains equally well to the grammar in the form of non-content items such as prepositions and articles or in the form of complex syntactic constructions - the idea that paradigmatic variation at the level of speech sounds should be excluded from the category of potentially meaningful items and relegated to pure form, to builders of meaning in the structuralist vein, was of course a disheartening thought.

Putting theory to the test, an empirical study on lectal acquisition was conducted in 2007. We know from research on accent-based speaker identification in adults that lectal recognition and classification is not only fast (Purnell, Idsardi and Baugh 1999) but also quite precise (van Bezooijen and Gooskens 1999) – but how does it work in young children? In Cognitive Linguistics it is assumed that language acquisition is a bottom-up process guided by linguistic experience. Input to language acquisition is an encounter with actual expressions and generalizations are made over usage-based events. In line with these principles we may also assume that knowledge of language *varieties* is experientially grounded and that lectal schemata emerge in terms of abstractions over usage-based events. So when do children begin to pay attention to not just *what* is said, but *how* it is said? If we assume that phonetic detail is not discarded but stored and put to constructive uses, how precise is lectal recognition across different age groups and what are the crucial stages in terms of emergence, or consolidation, of experientially grounded knowledge?

The interest in the explanatory dimension (*how* do children acquire the knowledge that allow them to identify accents as effectively as they do?) arose from a firm conviction that passive competence obviously does exist. To capture both dimensions in the design of the study we identified the following research questions:

- When do children begin to pay attention not only to what is said but also to how it is said?
- When do children begin to systematically store information about subtle differences in pronunciation, generalizing over usage-based events to form lectal categories?
- How accurate is speech perception at different intervals of age?
- How precise is it at different taxonomic levels?
- Which linguistic features and dimensions allow children to proceed to correct identification?
- Which factors – apart from age – have an influence on awareness?

In order to address these questions the two experiments described in this chapter were conducted in spring 2007. Adopting a usage-based approach to linguistic stereotype formation we predicted that structure would be determined by use and that purely formal characteristics (such as phonetic salience and contrast) would have fewer effects on correct identification than relative social salience (such as social stereotyping).

2.   Experiment 1: Identification of L1 accents

This experiment aimed at assessing the degree of correct identification of regional varieties in peninsular Spain by young native children across different age groups. As identification can take place at various levels within a taxonomic hierarchy of linguistic varieties, ranging from large-scale categorizations such as northern or southern speech, medium-scale categories such as Catalan or Andalusian and at a scale of increasingly more specific levels of abstraction, we were interested in eliciting not only the *degree* but also the level of *specificity* of the (correct or incorrect) identification on the part of the children. A suitable design to capture both dimensions thus had to be set up.

2.1. Method

Varieties from four regional Spanish dialects were assessed in the experiment: Galicia, Madrid, Andalucía and Gran Canarias. For each type two tokens were represented so as to test systematically for type-token relationships. Except for Galicia the tokens were from identical dialect areas: two speakers from Santa Cruz de Tenerife represented Canarias, two speakers from Seville represented Andalucía and two speakers from metropolitan Madrid represented the community of Madrid (the default case in the contemporary sociolinguistic situation in Spain). In the case of Galicia, one speaker was from the province of Lugo and the other from the province of Pontevedra. The stimuli consisted of digital recordings of a text read by speakers born and raised in the area in question. Speakers suspecting that the purpose of the recording was to retrieve a sample of their accent were discarded. The eight speakers selected, aged between 23 and 42, had all been born and raised in the respective provinces and spoke with the corresponding local accent.

The four accents were not chosen randomly. Three of the accents (Galicia, Madrid and Andalucía) are sufficiently geographically and linguistically different from one another to constitute a representative sample of accents that convey a cluster of phonetic features ranging high on perceptual salience. In the case of the Canarian Islands, diachronic reasons apart and specifically with the purpose of an analysis of folk perception in mind, the accent may best be described in terms of a subcategorization of speech from Andalucía. The distinctive features are thus less salient with respect to southern speech in general and presumably not easily identified as a speech pattern stemming specifically from the Canarian Islands. Furthermore, the four accents range from high to low on a scale of relative salience of social stereotyping. Table 1 captures these dimensions.

Table 1. Degrees of social stereotyping and formal distinctiveness

Accents	Social stereotyping	Linguistic distinctiveness
Andalucía (Sevilla x 2)	high	high
Canarias (Santa Cruz deTenerife x 2)	low	high
Galicia (Pontevedra/Lugo)	low	fairly low
Madrid (Madrid x 2)	(default)	(default)

Speech patterns stemming from Seville are highly distinctive in terms of formal features when compared with the speech patterns which are typical of Madrid (the default, standard accent) and highly stereotyped in the media: socially stereotyped characters in films and TV series frequently speak with accents from Andalucía. Accents from Galicia are much less stereotyped than speech from Andalucía and linguistically less distinctive, too, when compared to the speech patterns of Madrid. At the level of accent, i.e. when the grammar is excluded, differences lie more in intonation than in phonetics and phonology (cf. section 4). A Canarian accent, in turn, rates almost as high on formal distinctiveness as speech from Andalucía, but as low on social stereotyping as Galicia. By feeding these variables into the stimuli predictions may now be made regarding the question of whether formal distinctiveness or the degree of social stereotyping influences the degree of correct identification the most.

The location of the four accents is shown on the map in Figure 1. The map was distributed to the listeners together with the response forms and the questionnaire.

Figure 1. Map of Spain showing location of Autonomous Communities. The Canarian Islands are southwest with respect to the peninsula, off the west coast of Africa

A total of 150 children distributed across 3 age groups participated as listeners: 50 children born in 2000 (aged 6-7 in spring 2007 when the experiment took place), 50 born in 1998 (aged 8-9) and 50 born in 1994 (aged 12-13). The test subjects lived in a municipality approximately 30 kilometers from metropolitan Madrid, attended local schools in the area and most of them had been born and raised in the municipality.

The speech fragments to which the children were exposed were based on a short text read in Standard Spanish so that the levels of syntax, lexis and morphology were held constant. In turn, the fragment was highly cued for variation at the level of phonetics (e.g. by incorporating lots of instances of /s/, /r/ and /θ/ in initial, checked and final position) and intonation (e.g. by including a variety of different mood structures which in Spanish accents trigger different intonation patterns). The fragment simulated an in-

formal message left on an answering machine (e.g. by a close friend, spouse or relative). The fragment, which topic-wisely discussed practical affairs such as buying presents, going to the cinema and having dinner afterwards, was constituted by the following brief text:

> ¿Que tal? Oye, te llamo para decirte que no hace falta que cojas un regalo para Juan. Ya le he comprado los barcos que vimos el jueves. El rojo y el azul. Creo que todo iría más rápido si tú te haces cargo de lo otro. Te veo esta noche en el cine. Y por favor, esta vez no te olvides de sacar las entradas. Me apetece cenar después. ¿Crees que vas a tener tiempo? Bueno, ¡hasta entonces!

The duration of the speech samples was below 30 seconds for all eight speakers. The stimuli were presented to the children in random order:

(1) Andalucía (Seville, female)
(2) Galicia (Lugo, female)
(3) Community of Madrid (metropolitan Madrid, male)
(4) Andalucía (Seville, male)
(5) Galicia (Pontevedra, male)
(6) Gran Canarias (Santa Cruz de Tenerife, male)
(7) Community of Madrid (metropolitan Madrid, male)
(8) Gran Canarias (Santa Cruz de Tenerife, male)

Both male and female speakers participated, but gender was not believed to have a bearing on the results and was not treated as an independent variable.

The eight speech samples were successively located by the test subjects with respect to a three-level response form so as to elicit not only paradigmatic choices among regions (e.g. Asturias, Extremadura, Valencia) but also levels of specificity: the children were told to provide one answer only at a level at which they felt "safe" about the response, and if in doubt at the highest level of certainty. The response form, as shown in Table 1 below, did not reflect dialect areas and dialect continua as a linguist would depict them, but rather geographical and socio-political constructs at different levels of granularity: frequent and familiar place names readily evoked, or so we assume, by means of a metonymic association between place of origin and speech style when a stretch of unclassified speech is correlated with a series of prototypically organized models, or linguistic stereotypes, on the basis of relative similarity.

Table 2. Three-level response form distributed to test subjects

Area	Community	Province
north	Galicia	La Coruña / Lugo / Orense / Pontevedra
centre	Asturias	Asturias
south	Cantabria	Cantabria
east	Pais Vasco	Álava / Guipúzcoa / Vizcaya
west	Navarra	Navarra
	Aragón	Huesca / Teruel / Zaragoza
	Cataluña	Barcelona / Gerona / Lérida / Tarragona
	La Rioja	La Rioja
	Castilla y León	Ávila / Burgos / León / Palencia / Salamanca / Segovia / Soria / Valladolid / Zamora
	Comunidad de Madrid	Madrid
	Extremadura	Badajoz / Cáceres
	Castilla la Mancha	Albacete / Ciudad Real / Cuenca / Guadalajara/ Toledo
	Comunidad Valenciana	Alicante / Castellón / Valencia
	Andalucía	Almería / Cádiz / Córdoba / Granada / Huelva/ Jaén / Málaga / Sevilla
	Murcia	Murcia
	Islas Baleares	Islas Baleares
	Islas Canarias	Las Palmas / Santa Cruz de Tenerife

The widest possible level was as abstract as the cardinal points. Many of the 6-7 year-olds (who undertook the task with much seriousness, proud to be asked about their skills and to participate in an experiment), opted for these general choices whenever they felt uncertain about more specific possibilities. At the second level of specificity came the 17 Spanish autonomous communities, and finally, at the most difficult level, the 51 provinces that constitute the 17 communities. Needless to say, expectations for this level were low for all three age groups, but predictions were still that a gradual construal of emerging patterns even at such a refined level of abstraction would be reflected in the data.

The response form contained the following three questions for each of the eight speakers:

Speaker nº 1:
Origin: _____
What did you observe in his/her pronunciation that makes you think the speaker is from this area?
What other people or characters do you know who speak this way?

The procedure was as follows. The children were asked if they were willing to participate in a scientific experiment with the purpose of testing their ability to determine the origin of an unknown speaker. They were requested to take the task seriously, listen carefully and respond with rigor. The response forms, the maps and the questionnaires were distributed and instructions given as to how to fill them in. Then the speech samples were played in the order specified above, and after each speaker the panel of listeners answered the three questions posed regarding origin, significant linguistic features and social associations. Afterwards experiment 2 on the recognition of L2 accents was carried out (see section 3 for details) and finally the additional questionnaire was answered by two out of the three age groups (cf. section 4 for details).

2.2. Results and discussion

In this subsection we provide and discuss the results that stem from the first question in the response form: which is the origin of this speaker? The open questions in the response form will be dealt with separately in section 4 together with data from the questionnaire. The research questions addressed in this section thus involve:

·   When do children begin to pay attention not only to what is said but also to how it is said?
·   When do children begin to systematically store information about subtle differences in pronunciation, generalizing over usage-based events to form lectal categories?
·   How accurate is speech perception at different intervals of age?
·   How precise is it at different taxonomic levels?

The data were extracted from the 150 response forms and processed by a statistician. Table 3 shows the results at the most difficult level: that of province. As expected, results for this level of granularity were not statistically significant (except for a few cases as will be shown in Table 5 below) and no more than 3 out of the 8 fragments were correctly identified by any

of the three age groups. However, we may still observe, with satisfaction, how the percentages decrease (in the case of no single correct answer) and increase (in the case of one, two or three correct replies) in the right direction: while 68 percent of the 6-7 year-olds fail to correctly locate a single speech sample at the level of province, "bad" results for the 12-13 year-olds are down to 22 percent. Reversely, while 10 percent of the youngest children manage to identify two accents correctly, 34 percent of the oldest children are successful at this level of performance.

Table 3. Global results of Experiment 1 across age groups: percentages of correct answers per age group and province

		No. correct answers/age: province				
		0	1	2	3	
Age	6-7	34	10	5	1	50
		68.0%	20.0%	10.0%	2.0%	100.0%
	8-9	25	15	8	2	50
		50.0%	30.0%	16.0%	4.0%	100.0%
	12-13	11	19	17	3	50
		22.0%	38.0%	34.0%	6.0%	100.0%
Total		70	44	30	6	150
		46.7%	29.3%	20.0%	4.0%	100.0%

At the less fine-grained level of autonomous community, we do obtain significant results: see Table 4. The number of correctly identified accents is now up to 5: although none of the youngest children have more than 4 correct results, 18 percent of the 12-13 year-olds successfully locate 5 accents at this level of abstraction. Again the youngest children are still learning how to build up schemas of lectal varieties: 38 percent fail to locate any of the eight accents in a successful manner, but the figures for both the 8-9 year-olds (20 percent) and the 12-13 year-olds (2 percent) show that they have acquired such skills in just a few years and they are now much less at a loss. It is important to notice how systematically the scores decrease for the bad results and begin to increase as the number of correct replies grows: we witness how patterns in acquisitional competence emerge in systematic manners from the age of 6 to 12.

Table 4. Global results of Experiment 1 across age groups: percentages of correct answers per age group and autonomous community

		No. correct answers/age: community						Total
		0	1	2	3	4	5	
Age	6-7	19	16	8	5	2	0	50
		38.0%	32.0%	16.0%	10.0%	4.0%	.0%	100.0%
	8-9	10	12	10	9	5	4	50
		20.0%	24.0%	20.0%	18.0%	10.0%	8.0%	100.0%
	12-13	1	5	15	10	10	9	50
		2.0%	10.0%	30.0%	20.0%	20.0%	18.0%	100.0%
Total		30	33	33	24	17	13	150
		20.0%	22.0%	22.0%	16.0%	11.3%	8.7%	100.0%

Let us now, in Table 5, cast a glance at a series of tables that aim to capture these various dimensions in terms of a global overview. The first column enlists age groups and the eight accents (where e.g. Andalucía 1 and Andalucía 2 represent the two tokens per type). The next three columns reflect cumulative percentages of correct answers per area, autonomous community and province. The results are cumulative because the scores of province are included in the scores of community and these in the results for area: if a child responded correctly at the level of province we assume that the wider levels of community and area were correctly identified, too.

The last three columns show the various levels of statistical significance (analyses based on Chi-squares): the results that proved to be positively significant for the 6-7 year-olds are marked in bold. (Results that on first sight are identical to the ones marked in bold were also statistically significant, but in the *negative* sense: the scores were unusually low and thus stood out in a sense that on the other hand is still relevant for the purposes of the present study. When an accent rates very *low* on awareness we are obviously still interested in the processes behind such a lack of knowledge or attention. In this paper, however, we shall concentrate on the results that pertain to a *high* degree of awareness.) Let us first of all observe that type-token correspondences are respected: when Andalucía 1 is identified at levels above chance, so is Andalucía 2, and as we shall see in the next two tables, the type/token relationship is systematic for all three age groups. This is not a result without importance: it indicates that identification was not a random affair, but that patterns (or more technically speaking linguis-

tic stereotypes: models capturing the nature of prototypically organized lectal categories) are emerging that operate in the same manners for the same tokens. For this panel of young listeners the two Andalusian accents were identified correctly at the level or area (south) and community (Andalucía) and so were the scores for Madrid at all three levels. It must be borne in mind, however, that Madrid is a special case for at least three reasons: first, because it represents the default standard accent, omnipresent in the media. Second, because it is the predominant accent spoken in the children's own geographical surroundings, and third, because the autonomous community of Madrid is composed of one province only: the province of Madrid. Correct results at the level of province were only attributed when the children explicitly stated the term province in their reply, but we still expect the good results at the most fine-grained level to be influenced by the correspondence between hypernym and hyponym.

Table 5. Global results of Experiment 1 across age groups and response levels: age group 6-7

Age/Location	Area	Com	Prov	Area	Com	Prov
6-7				0.2 (1/5)	0.0588 (1/17)	0.02 (1/50)
Andalucía 1	60	32	4	**0.0000**	**0.0000**	0.3124
Galicia 1	6	0	0	0.0133	0.0772	0.3124
Madrid 1	30	16	16	**0.0771**	**0.0024**	**0.0000**
Tenerife 1	10	4	0	0.0771	0.5720	0.3124
Andalucía 2	32	24	4	**0.0339**	**0.0000**	0.3124
Galicia 2	8	6	0	0.0339	0.9712	0.3124
Madrid 2	36	22	22	**0.0047**	**0.0000**	**0.0000**
Tenerife 2	10	6	0	0.0771	0.9712	0.3124
8-9						
Andalucía 1	62	52	4	0.0000	0.0000	0.3124
Galicia 1	6	4	0	0.0133	0.5720	0.3124
Madrid 1	62	52	40	0.0000	0.0000	0.0000
Tenerife 1	6	0	0	0.0133	0.0772	0.3124
Andalucía 2	34	32	0	0.0133	0.0000	0.3124
Galicia 2	20	14	0	1.0000	0.0147	0.3124
Madrid 2	48	42	30	0.0000	0.0000	0.0000
Tenerife 2	2	2	0	0.0015	0.2435	0.3124

12-13						
Andalucía 1	86	82	16	0.0000	0.0000	0.0000
Galicia 1	18	14	0	0.7237	0.9712	0.3124
Madrid 1	70	52	50	0.0000	0.0000	0.0000
Tenerife 1	26	6	0	0.2888	0.9712	0.3124
Andalucía 2	56	50	20	0.0000	0.0000	0.0000
Galicia 2	56	42	0	0.0000	0.0000	0.3124
Madrid 2	62	40	36	0.0000	0.0000	0.0000
Tenerife 2	38	14	2	0.0015	0.0147	1.0000

Will the results for the 8-9 year-olds resemble the performance of the youngest children with respect to types identified and levels of significance? In Table 6 we add the scores for the next age group.

Table 6. Global results of Experiment 1 across age groups and response levels: age group 8-9

Age/Location	Area	Com	Prov	Area	Com	Prov
6-7				0.2 (1/5)	0.0588 (1/17)	0.02 (1/50)
Andalucía 1	60	32	4	**0.0000**	**0.0000**	0.3124
Galicia 1	6	0	0	0.0133	0.0772	0.3124
Madrid 1	30	16	16	**0.0771**	**0.0024**	**0.0000**
Tenerife 1	10	4	0	0.0771	0.5720	0.3124
Andalucía 2	32	24	4	**0.0339**	**0.0000**	0.3124
Galicia 2	8	6	0	0.0339	0.9712	0.3124
Madrid 2	36	22	22	**0.0047**	**0.0000**	**0.0000**
Tenerife 2	10	6	0	0.0771	0.9712	0.3124
8-9						
Andalucía 1	62	52	4	**0.0000**	**0.0000**	0.3124
Galicia 1	6	4	0	0.0133	0.5720	0.3124
Madrid 1	62	52	40	**0.0000**	**0.0000**	**0.0000**
Tenerife 1	6	0	0	0.0133	0.0772	0.3124
Andalucía 2	34	32	0	**0.0133**	**0.0000**	0.3124
Galicia 2	20	14	0	1.0000	**0.0147**	0.3124
Madrid 2	48	42	30	**0.0000**	**0.0000**	**0.0000**
Tenerife 2	2	2	0	0.0015	0.2435	0.3124

12-13						
Andalucía 1	86	82	16	0.0000	0.0000	0.0000
Galicia 1	18	14	0	0.7237	0.9712	0.3124
Madrid 1	70	52	50	0.0000	0.0000	0.0000
Tenerife 1	26	6	0	0.2888	0.9712	0.3124
Andalucía 2	56	50	20	0.0000	0.0000	0.0000
Galicia 2	56	42	0	0.0000	0.0000	0.3124
Madrid 2	62	40	36	0.0000	0.0000	0.0000
Tenerife 2	38	14	2	0.0015	0.0147	1.0000

As in the case of the results for the 6-7 year-olds we have marked the re-
sults that were statistically significant in bold. We observe that the very
same categorizations that were located at levels above chance in the case of
the 6-7 year-olds are also identified correctly by the 8-9 year-olds. And
now one more accent is added to the list of significant scores: Galicia 2 at
the level of autonomous community (marked in grey). That the same results
were obtained in the case of the second age group and that one new accent
is added to the picture speaks of a systematic development. Had the child-
ren at this age identified other accents than the ones located by the young-
est listeners we would be in need of different parameters than the ones we
brought into the picture in the context of the investigation.

    Would the results of the oldest age group confirm the tendency observed
in the previous two tables? In Table 7 the results that are statistically sig-
nificant are likewise highlighted in bold and significant values that are new
for this age group are again marked in grey.

Table 7. Global results of Experiment 1 across age groups and response levels: age
        group 12-13

Age/Location	Area	Com	Prov	Area	Com	Prov
6-7				0.2 (1/5)	0.0588 (1/17)	0.02 (1/50)
Andalucía 1	60	32	4	**0.0000**	**0.0000**	0.3124
Galicia 1	6	0	0	0.0133	0.0772	0.3124
Madrid 1	30	16	16	**0.0771**	**0.0024**	**0.0000**
Tenerife 1	10	4	0	0.0771	0.5720	0.3124
Andalucía 2	32	24	4	**0.0339**	**0.0000**	0.3124

Galicia 2	8	6	0	0.0339	0.9712	0.3124
Madrid 2	36	22	22	**0.0047**	**0.0000**	**0.0000**
Tenerife 2	10	6	0	0.0771	0.9712	0.3124
8-9						
Andalucía 1	62	52	4	**0.0000**	**0.0000**	0.3124
Galicia 1	6	4	0	0.0133	0.5720	0.3124
Madrid 1	62	52	40	**0.0000**	**0.0000**	**0.0000**
Tenerife 1	6	0	0	0.0133	0.0772	0.3124
Andalucía 2	34	32	0	**0.0133**	**0.0000**	0.3124
Galicia 2	20	14	0	1.0000	**0.0147**	0.3124
Madrid 2	48	42	30	**0.0000**	**0.0000**	**0.0000**
Tenerife 2	2	2	0	0.0015	0.2435	0.3124
12-13						
Andalucía 1	86	82	16	**0.0000**	**0.0000**	**0.0000**
Galicia 1	18	14	0	0.7237	**0.0147**	0.3124
Madrid 1	70	52	50	**0.0000**	**0.0000**	**0.0000**
Tenerife 1	26	6	0	0.2888	0.9712	0.3124
Andalucía 2	56	50	20	**0.0000**	**0.0000**	**0.0000**
Galicia 2	56	42	0	**0.0000**	**0.0000**	0.3124
Madrid 2	62	40	36	**0.0000**	**0.0000**	**0.0000**
Tenerife 2	38	14	2	**0.0015**	**0.0147**	1.0000

It is beyond doubt interesting to see how the same accents that were identified at levels above change by the two first age groups recur and how new accents and new levels of specificity are added to the picture: both tokens of the Andalusian accent are now also identified at the level of province (Seville), Galicia 1 now appears and so does Tenerife 2. The image we obtain of the abilities of the pre-adolescents regarding receptive competence of lectal variation is quite complete and clearly represents a growing degree of awareness in comparison with the younger children.

3.   Experiment 2: Identification of L2 accents

The second experiment, conducted with the same groups of children as a continuation of experiment 1, aimed to assess the degree of correct identifi-

cation of L2 accents. This test was not less "tough" than the first: to what extent can a child discern the origin of a non-native speaker of peninsular Spanish on the basis of transfer from the speaker's mother tongue (or mother accent in the case of Latin American speakers)?

3.1. Method

To assess this question, 7 additional instances of the same text as in experiment 1 were recorded. This time only one of the speakers was a native from peninsular Spain. The rest were L2 speakers with a medium degree of competence so as to allow for transfer from the speaker's L1 language and L1 speakers of a Latin American variety. The origin and order of the samples were as follows:

(1) United States of America (Washington D.C., female)
(2) Argentina (Buenos Aires, male)
(3) Germany (Nürnberg, female)
(4) Spain (Zaragoza, female)
(5) Mexico (Puebla de Zaragoza, female)
(6) England (Manchester, female)
(7) France (Reims, female)

The children were provided with a multiple-choice form:

Indicate with numbers (1-7) where these speakers are from:
☐   Spain
☐   Argentina
☐   Mexico
☐   France
☐   Germany
☐   The United States of America
☐   England

Furthermore, for each of the 7 accents the children were asked the same two open questions as in Experiment 1:

Speaker nº 1:
Origin: _____
What did you observe in his/her pronunciation that makes you think the speaker is from this area?
What other people or characters do you know who speak this way?

3.2. Results and discussion

From a global perspective, the three age groups perfomed as in Table 8. The table represents the percentage of correct answers across the three age groups for the seven speech samples. Let us note once again how successful performance increases systematically and significantly: 28 percent of the 6-7 year-olds were only able to recognize one accent correctly (and as we shall see in brief in the vast majority of the cases this was their own native accent: peninsular Spanish), but this low score is down to 10 percent in the case of the 8-9 year-olds and 0 percent for the 12-13 year-olds. In turn, none of the youngest children have 7 correct answers, but 8 percent of the teenagers do. The highest percentage of correct answers falls in the upper middle of the range: almost half of the 12-13 year-olds (24 out of 50) have 4 correct answers out of 7 and almost one quarter (24 percent) have 5. We may thus safely conclude that awareness of non-native lectal varieties increase systematically in children in a time span of six years, reaching a fairly high level of precision in early adolescence.

Table 8. Global results of Experiment 2 across age groups

		Number of correct answers/age group: all 7 countries						Total
		1	2	3	4	5	7	
Age	6-7	14	18	7	5	6	0	50
		28.0%	36.0%	14.0%	10.0%	12.0%	.0%	100.0%
	8-9	5	18	10	8	7	2	50
		10.0%	36.0%	20.0%	16.0%	14.0%	4.0%	100.0%
	12-13	0	3	7	24	12	4	50
		.0%	6.0%	14.0%	48.0%	24.0%	8.0%	100.0%
Total		19	39	24	37	25	6	150
		12.7%	26.0%	16.0%	24.7%	16.7%	4.0%	100.0%

Let us now cast a deeper glance at the data. Which languages or accents were more readily recognized than others? Table 9 illustrates the percentages for the three L1 accents (Argentinean, Mexican and peninsular Spanish).The success rate of the 6-7 year-olds has now increased to 12 percent and the 12-13 year-olds perform with much accuracy on this dimension: 41 out of 50 test subjects knew who spoke with an Argentinean, a Mexican or a peninsular Spanish accent.

Table 9. Results of L1 accents across age groups

		Number of correct answers/age group: L1 accents				Total
		0	1	2	3	
Age	6-7	4	25	15	6	50
		8.0%	50.0%	30.0%	12.0%	100.0%
	8-9	0	10	27	13	50
		.0%	20.0%	54.0%	26.0%	100.0%
	12-13	1	1	7	41	50
		2.0%	2.0%	14.0%	82.0%	100.0%
Total		5	36	49	60	150
		3.3%	24.0%	32.7%	40.0%	100.0%

Table 10. Results for peninsular Spanish for age group 6-7

6-7	Frequency	Percentage
Correct	43	86.0
Germany	1	2.0
France	1	2.0
England	4	8.0
Mexico	1	2.0
Total	50	100.0

Table 11. Results for peninsular Spanish for age group 8-9

8-9	Frequency	Percentage
Correct	50	100.0

Table 12. Results for peninsular Spanish for age group 12-13

12-13	Frequency	Percentage
Correct	49	98.0
Argentina	1	2.0
Total	50	100.0

Table 13. Results for Argentinean for age group 6-7

6-7	Frequency	Percentage
Correct	18	36.0
Germany	6	12.0
France	4	8.0
England	12	24.0
Mexico	7	14.0
USA	3	6.0
Total	50	100.0

Table 14. Results for Argentinean for age group 8-9

8-9	Frequency	Percentage
Correct	34	68.0
Germany	4	8.0
France	5	10.0
England	1	2.0
Mexico	5	10.0
USA	1	2.0
Total	50	100.0

Table 15. Results for Argentinean for age group 12-13

12-13	Frequency	Percentage
Correct	48	96.0
France	1	2.0
Mexico	1	2.0
Total	50	100.0

Pushing the question further, how readily will children know when a speaker is native or a foreigner? Tables 10 to 11 list the results for the peninsular Spanish speaker across the three age groups. 86 percent of the youngest children knew when a speaker is not a foreigner. By keeping track of the wrong choices we can also tell where the children who thought the speaker had a different origin believed the speaker was actually from: in the case of the youngest children, the peninsular accent was erroneously attri-

buted to Germany, France, England and Mexico by a total of 7 out of 50 listeners. In the case of the 8-9 year-olds, there is no doubt at all about the condition of nativeness: see Table 11. If 100 percent of the 8-9 year-olds knew who spoke with a peninsular accent and who were foreign, the implications for social stereotyping and processes of social exclusion (cf. Purnell, Idsardi and Baugh 1999) are obvious, even in young children. The results for the 12-13 year-olds were as in Table 12. One out of 50 children took the peninsular Spanish speaker to be Argentinean, but the remaining 49 correctly identified the accent as peninsular Spanish. There can thus be little doubt about the degree of awareness of lectal variation at the most fundamental level: children soon know when someone speaks with a foreign accent and when not.

How, then, were the near-native accents evaluated? In the case of Argentina, the results were as in Table 13. 18 out of 50 of the youngest identifiers managed to correctly locate this accent as Argentinean, but as many as 12 incorrectly identified it as a British English accent. In the case of the 8-9 year-olds (Table 14), the success rate increased by almost 50 percent. In the case of the 12-13 year-olds (Table 15), the results show an even higher degree of precision. Only two children out of 50 attributed the Argentinean accent to other identities (French and Mexican, respectively).

Further, how did the L2 accents fare on a fine-grained basis in comparison with the native accents? One would expect an accent so well-known amongst youngsters as American English (due to the influence of pop songs, popular TV series etc) to be readily identified as such. However (see section 4 for a discussion) this was far from the case: see Table 16. The confusion is generalized: the youngest children attribute the American accent almost randomly to all sorts of varieties, with their native peninsular Spanish, England and Germany as minor exceptions. Not even by adding the results of England to those of the USA (both being subcategories of English) do the correct results surpass the scores of France.

The 8-9 year-olds fared only slightly better. If adding the scores for the USA and England one obtains a percentage of correct results of 30 percent – but even so the scores for France and Argentina together sum up 40 percent of the (in these cases incorrect) evaluations: see Table 17. In the case of the 12-13 year-olds the percentage of English accents, however, is doubled. 60 percent of the adolescents now identify the accent as foreign and the language behind the accent as English. Whether the accent is American or British there is much less agreement: see the results in Table 18.

That the global results for France, Germany and England present the same degree of lack of precision is clear from the summary in Table 19.

Table 16. Results for the USA. Age group 6-7

6-7	Frequency	Percentage
Correct	8	16.0
Germany	6	12.0
Argentina	9	18.0
Spain	2	4.0
France	12	24.0
England	4	8.0
Mexico	9	18.0
Total	50	100.0

Table 17. Results for the USA. Age group 8-9

8-9	Frequency	Percentage
Correct	7	14.0
Germany	6	12.0
Argentina	9	18.0
France	11	22.0
England	8	16.0
Mexico	9	18.0
Total	50	100.0

Table 18. Results for the USA. Age group 12-13

12-13	Frequency	Percentage
Correct	14	28.0
Germany	9	18.0
Argentina	1	2.0
France	5	10.0
England	16	32.0
Mexico	5	10.0
Total	50	100.0

Table 19. Correct results in percentages for L1/L2 experiment sorted per variety, age and success rate

Age	Spain	Argentina	Mexico	France	Germany	England	USA
6-7	86	36	24	18	40	22	16
8-9	100	68	38	22	38	24	14
12-13	98	96	82	54	36	28	28

When ordered in terms of the success rate we observe that native peninsular Spanish comes in as a clear winner as far as correct identification is concerned. The non-native L1 accents (Argentinean and Mexican) are less successful for the youngest children, but achieve results similar to the native mother tongue in the case of the oldest age group. For the 12-13 year-olds France leads the group of the L2 accents, but in general the results are worse than the L2 scores, and there is hardly any difference between the languages, nor is the increase across age groups as significant as with the L1 languages. Germany is furthermore the only language in which the rate of correct identification is reversed: while the 6-7 year olds identify it correctly as German, surprisingly the older age groups both have a lower score. Possible reasons why will be discussed in the next section. All in all, we may for now conclude that awareness of lectal differences systematically increases even in the case of L2 accents, though to a minor degree when compared to L1 varieties.

4.   The questionnaire: the experiential basis of lectal competence

So far we have concentrated on the descriptive dimension only: what do children know about lectal varieties and when do they acquire the knowledge? We by now know that lectal awareness is an early acquisition and that increase is systematic across ages and type-token relationships. It is now time to address the explanatory dimension, as well: where does this awareness stem from? Why these accents, and in this order? Is it exclusively the result of the growth of cognitive capacities? How do they build up the schemas and acquire the ability? In this section we thus address the last two research questions identified in section 1:

· Which linguistic features and dimensions allow the children to proceed to correct identification?
· Which factors – apart from age – have an influence on awareness?

Let us first of all mention the fact that the questionnaire designed to elicit data related to the second question was distributed to the 8-9 and the 12-13 year-olds but not to the children aged 6-7. In a pilot study conducted previously in order to specify factors such as the optimal length of the stimuli, the most convenient age groups and the type of questions to be incorporated in the final design, it became clear that a questionnaire would only be a useful instrument for the age groups in which children have command of all the notions involved in a given question and are able to provide reliable answers to the questions posed. For a mixture of cognitive and practical reasons, all but the youngest age group received the questionnaire. In the pilot study questions such as "How many countries have you visited?", regions were often provided as samples of countries (e.g 3: France, Valencia and Portugal). Note that lack of adequate notions of geographical and political categorisations did not prevent the children from carrying out the experiments in satisfactory ways. In order to associate a stimulus with the name of the region or place evoked no geographical or political knowledge is required. In the same vein, when 6-year-olds were asked to quantify the number of hours they watch TV in the course of a normal week, vague or inaccurate results were frequently yielded. To stipulate the duration of an event in terms of units such as hours was an easy task for the 8-9 year-olds, but not for the youngest children. The 6-7 year olds were thus discarded as informants for this part of the study (but did answer the two open questions formulated in the response form regarding salience of linguistic structure and social associations as there were no impediments to comprehension in these cases). The questionnaire gathered data on the following dimensions.

1. Age
2. Number of countries visited so far
3. Nature of countries visited so far
4. Places frequently visited during holidays or weekends
5. Places of residence
6. Origin of mother
7. Origin of father
8. Origin of maternal/paternal grandparents and current place of residence
9. Origin of caretakers/teachers/friends/neighbours
10. Number of hours spent watching TV/week
11. Number and nature of accents that you know how to imitate

What, then, do these data tell us about the explanatory dimension of lectal acquisition? In order to try to throw light on the possible predictors of the success rate for the results obtained for the three age groups in question, 5 measures of objective and subjective distances were identified:

1. Objective distance I: number of distinctive linguistic features with respect to a reference value (Standard Spanish)
2. Objective distance II: number of exclusive features (i.e. non–shared types)
3. Objective distance III: number of distinctive features (types) with a high token frequency
4. Subjective distance I: number of distinctive features the subjects were aware of
5. Subjective distance II: number of familiar speakers and socially stereotyped characters

Let us now address each of these measures in more detail and relate the different dimensions to the data obtained. The results are currently being subjected to a multivariate regression analysis, which means that in what follows we will rely on rough counts only, but the findings are still telling.

4.1. Linguistic features as an objective distance

It has been argued from the perspective of cognitive phonology (e.g. Taylor 1990, 1995, 2002) that linguistic distance of phonetic features plays a role in the construal of prototypically conceived phonemic categories. Centrality of phonemic categories is naturally ascribed to certain allophones because of structural properties inherent in the language:

> […] the putative central member of /t/ - say, the voiceless aspirated alveolar plosive – enters into a number of highly salient perceptual and articulatory contrasts with the putative central members of neighbouring categories, such as the unaspirated alveolar plosive of /d/, the voiceless aspirated velar plosive of /k/, and so on. (Taylor 1995: 228)

On the other hand it has been suggested (Kristiansen 2001, 2003, 2006, 2008) that intraphonemic and transphonemic distinctiveness is exploited by language users in order to navigate in the social world: that unique clusters of perceptually salient linguistic features are socially distinctive and operate as cognitive reference points, or linguistic stereotypes, that effectively evoke their social counterparts: social categorizations and social stereo-

types. In other words, if lectal varieties form prototype categories, the central images will consist of a unique combination of relatively shared linguistic features.

It is thus quite obvious to raise the following question: would it be the case that the farther away from the speech of Madrid (as a default reference value for the listeners) a variety is linguistically speaking, the easier it is to recognize the accent in question? I.e. at more phonetic distance, better identification?

Table 20. Main linguistic differences in speech fragments when contrasted with a default variety: qualitative overview

Feature description	And	Gal	Can	Arg	Mex	Fra	Ger	BrE	AmE
/r/ as uvular fricative		✓				✓	✓	✓	
marked tone unit clause final		✓							
marked tone unit interrogative mood		✓		✓					
/Ɵ/ as [s]	✓		✓	✓	✓	✓	✓		
final and checked /s/ dropped	✓		✓						
final /r/ dropped	✓								
intervocalic /s/ as marked aspiration			✓						
isolated vowel change (apatece)	✓		✓						
final /l/ as [ɫ]								✓	✓
/r/ as back uvular									✓
/y/ as [ʒ] or [ʃ]				✓					
/d/ as alveolar						✓	✓	✓	✓
marked tone unit exclamative mood					✓				
aspirated initial /t/								✓	
vowel changes: close e/o, [æ] for [a]						✓			

In Table 20 we present the perceptually most salient linguistic differences - i.e. we consider only those contrasts that are immediately salient and audible at a fairly general level - that which characterized the speech fragments implemented in the test with respect to the default accent. As we would expect, most of the features are shared with other accents, but the combinations as such are unique and thus socially distinctive. If we now quantify the differences at this quite general, coarse level of description we observe that the differences numerically speaking range between 2 and 5, as indicated in Table 21. As the accents appear in order of correct identification,

we conclude that the success rate cannot be predicted on the basis of objective linguistic distance alone. There are no purely linguistic reasons why British English, with 4 main contrastive features, should be identified by the children less readily than German or Mexican, with fewer salient contrasts.

Table 21. Main linguistic differences in speech fragments when contrasted with a default variety: quantitative overview

	Distinctive features	Exclusive features	Frequent features	Linguistic awareness	Social awareness
Andalucía	**4**				
Galicia	**3**				
Canarias	**4**				
Argentina	**3**				
Mexico	**2**				
France	**4**				
Germany	**3**				
Br. Eng.	**4**				
Am. Eng.	**3**				

4.2. Exclusive linguistic features as an objective distance

Our next candidate for a good predictor of the success rate is linguistic exclusiveness (cf. Kristiansen 2003, 2008). Structural features that are not shared with other accents are likely to rate high on perceptual salience and social awareness. Let us recall in this respect that Labov once described degrees of social awareness in terms of a gradient from social stereotypes to markers, and finally indicators. Furthermore, as pointed out by Nunberg (1978) in his book on metonymic reference, the ideal signifier is perceptually distinctive and functionally exclusive: forms that relate to several referents *characterize*, but forms that relate exclusively to a single referent *identify*. An accent with several exclusive structural components should thus in theory be identified with more ease than an accent with one, or no, exclusive features. In Table 22 we have marked the features that occur in only one of the speech fragments under scrutiny. We observe that all the accents, except for German, exhibit one feature which is not shared by any of the other lects: see Table 23.

Table 22. Exclusive linguistic differences when contrasted with default variety: qualitative overview

Feature description	And	Gal	Can	Arg	Mex	Fra	Ger	BrE	AmE
/r/ as uvular fricative									
marked tone unit clause final		✓							
marked tone unit interrogative mood									
/Ɵ/ as [s]									
final and checked /s/ dropped									
final /r/ dropped	✓								
intervocalic /s/ as marked aspiration			✓						
isolated vowel change (apatece)									
final /l/ as [ɫ]									
/r/ as back uvular									✓
/y/ as [ʒ] or [ʃ]				✓					
/d/ as alveolar									
marked tone unit exclamative mood					✓				
aspirated initial /t/								✓	
vowel changes: close e/o, [æ] for [a]						✓			

Table 23. Exclusive linguistic differences when contrasted with default variety: quantitative overview

	Distinctive features	Exclusive features	Frequent features	Linguistic awareness	Social awareness
Andalucía	4	**1**			
Galicia	3	**1**			
Canarias	4	**1**			
Argentina	3	**1**			
Mexico	2	**1**			
France	4	**1**			
Germany	3	**0**			
Br. Eng.	4	**1**			
Am. Eng.	3	**1**			

Exclusiveness, then, does not explain the order of correct identification, either. What the lack of exclusive features in the German speech fragment might help to explain, though, is the fact that it is the only accent for which

the results worsened for the two oldest age groups. We suspect that lack of a clear reference point may well have contributed to hindering identification in age groups where lectal schemas in general have become more developed.

4.3. Frequency of distinctive features as an objective distance

If neither linguistic contractiveness nor exclusiveness can explain the differences in the results, would frequency of occurrence of distinctive features then play a role? Hearing a socially distinctive feature just a few times in a short fragment might yield vaguer results than if a feature is repeated many times. To account for this possibility we counted the distinctive linguistic features that occurred more than twice in the fragments. As can be seen in Table 24, however, the number of high frequency distinctive features is not a very helpful dimension. It is far from the case that the accents with the most high frequency features are identified first. If the knowledge is there, hearing just a few features a few times seems to be enough for the match between linguistic stereotype and social category to take place.

Table 24. Number of high frequency features

	Distinctive features	Exclusive features	Frequent features	Linguistic awareness	Social awareness
Andalucía	4	1	**2**		
Galicia	3	1	**2**		
Canarias	4	1	**2**		
Argentina	3	1	**1**		
Mexico	2	1	**1**		
France	4	1	**3**		
Germany	3	0	**3**		
Br. Eng.	4	1	**3**		
Am. Eng.	3	1	**3**		

4.4. Linguistic awareness as a subjective distance

If the reasons behind the success rate do not primarily lie in the language, it will rather be a matter of usage. Let us accordingly now turn to factors that pertain to the users and to patterns of language usage. To what extent are children aware of the features they make use of in order to evoke lectal schemata? For each speech fragment the 150 children were asked what they observed in the speaker's pronunciation that made them think he/she was from the area in question. There is no guarantee that the objective linguistic differences discussed above were actually perceived as such by the listeners, but the features they *did* perceive are likely to be highly salient and contrastive, and if lectal schemata arise in a bottom-up fashion, the more features reported on and the higher the level of awareness, the more entrenched a lectal categorisation should in theory be. So which features did they report on and how many were mentioned? Table 25 captures these two dimensions.

In total there were 32 references to the various instantiations of /s/ in the fragments from Andalucía (e.g. /s/ dropping, alternations between /s/ and /Ɵ/) which are indeed typical and frequently quoted features of this region. Two children however also observed that <j> as in <Juan> was pronounced in a more aspirated manner than the clearly fricative /x/ in the speech of Madrid. This is a feature which was not deemed to be perceptually salient enough to be included as an objective variant in Table 20 above, but a few of the children still captured the distinctiveness of the pronunciation. In the case of the Canarian islands, 14 children mention variants of /s/, 2 refer to the aspiration of /x/ and 1 child comments on /t/. For the accents from Galicia the children also mention two variables: 4 refer to different realizations of /s/ (such as the fact that the speakers *do* pronounce them) and there are as many as 10 references to differences in intonation patterns. This leads us to think that Galician speech was less objectively non-distinctive for the children than hypothesized in section 2 (cf. Table 1). Tone units seem to be perceived almost as readily as a missing allomorph. This is of course only in consonance with usage-based theories such as exemplar-based models (e.g. Bybee 2001, 2003, Pierrehumbert 2001) according to which linguistic representation is shaped by speakers´ memories of specific tokens of linguistic features. Phonetic detail is not discarded but stored in long-term memory, and so are subtle differences in intonation patterns.

Table 25. Subjective awareness of linguistic features: qualitative and quantitative overview (tokens)

Accent	Nature of type mentioned and number of token
Andalucía	32 /s/, 2 aspiration of /x/ in <j>
Galicia	4 /s/, 10 intonation patterns
Canarias	14 /s/, 2 /x/, 1 /t/
Argentina	11 / sibilants: /s/ and affricate <y>, 1 /r/, 1 intonation
Mexico	11 /s/, 2 /r/
France	9 /r/, 3 /s/, 1 <g>, 1 /t/
Germany	5 /r/, 1 <j>, 1 <g>
Br. Eng.	7 /r/, 3 /Ө/ ⇨/s/, 2 /t/, 3 /g/, 2 /o/
Am. Eng.	5 /r/, 2 /s/

Table 26. Subjective awareness of linguistic features: quantitative overview (types)

	Distinctive features	Exclusive features	Frequent features	Linguistic awareness	Social awareness
Andalucía	4	1	2	**3**	
Galicia	3	1	2	**2**	
Canarias	4	1	2	**3**	
Argentina	3	1	1	**3**	
Mexico	2	1	1	**2**	
France	4	1	3	**4**	
Germany	3	0	3	**3**	
Br. Eng.	4	1	3	**5**	
Am. Eng.	3	1	3	**2**	

In the case of British English, a total of 17 children comment on 5 different variables. In our ranking of successful identification we separated the results for British and American English, but it must not be forgotten that these are ultimately two varieties of the same language and that as we saw in Table 18, in the oldest age group 14 out of 50 identified American English correctly and 16 thought is was a British English accent. 60 percent thus knew that the lects were English. Out of the remaining 20 listeners, 9 children believed it was German and only a few subjects attributed the speech to a "southern" origin (France, Argentina, Mexico). It is thus only if we keep the English varieties separately that France and Germany fare slightly better in spite of there being fewer comments and fewer variables

enlisted for these accents. In Table 26 we specify the number of types quoted by the subjects.

4.5. Social awareness and familiarity as a subjective distance

If lectal categories are experientially grounded, based on real encounters with linguistic variants and variables in purposeful usage in meaningful contexts, the idea that social and linguistic awareness should increase with exposure to social and linguistic variety is a natural corollary.

To what extent would travelling to foreign countries for instance affect the success rate? We hypothesize that widely travelled children receive more exposure to variation and that this increase in exposure results in a higher degree of lectal awareness.

Table 27. Familiarity as a subjective distance: the effects of travelling on foreign accent identification

	Correlation correct answers in experiment 2/countries visited					Total
Total hits in test	2	3	4	5	7	
0 countries visited	1	2	6	2	2	13
% countries visited/hit	7.7%	15.4%	46.2%	15.4%	15.4%	100%
1-2 countries visited	1	3	9	7	0	20
% countries visited/hit	5.0%	15.0%	45.0%	35.0%	0.0%	100%
3 + countries visited	1	2	9	3	2	17
% countries visited/hit	5.9%	11.8%	52.9%	17.6%	11.8%	100%
Total countries visited	3	7	24	12	4	50
Total % countries visited/hit	6.0%	14.0%	48.0%	24.0%	8.0%	100%

Table 27 correlates the number of countries visited by the 12-13 year-olds with the results they obtained in the second experiment (foreign accents). Contrary to our expectations, however, the results do not improve as the result of exposure to cultures and languages in the form of travelling to foreign countries. The children who had visited more than 3 countries hardly improved their results with respect to the ones who had never been abroad. By way of example, in the case of the children who had 4 correct

results out of 7 (48 % of the 50 subjects), 6 subjects had never been abroad, 9 had been to one or two countries and 9 were more widely travelled. The children who obtained 5 correct answers out of 7 possible hits had been to no foreign country in 15.4 percent of the cases, to one or two destinations in 35 percent and only 17.6 had visited more than three countries. There seems to be no significant correlation between travelling and receptive lectal competence.

Table 28. Familiarity as a subjective distance: the effects of exposure to the media on L1 dialect identification at the level of autonomous community

		Correlation correct answers in experiment 1/ exposure to TV						Total
Total hits in test		0	1	2	3	4	5	
TV - 2		0	0	2	0	1	3	6
%		.0%	.0%	33.3%	.0%	16.7%	50.0%	100.0%
3-7		0	0	4	4	4	3	15
%		.0%	.0%	26.7%	26.7%	26.7%	20.0%	100.0%
8-14		0	3	3	1	3	2	12
%		.0%	25.0%	25.0%	8.3%	25.0%	16.7%	100.0%
15 +		1	2	6	5	2	1	17
%		5.9%	11.8%	35.3%	29.4%	11.8%	5.9%	100.0%
Total nº		1	5	15	10	10	9	50
Total %		2.0%	10.0%	30.0%	20.0%	20.0%	18.0%	100.0%

If competence does not stem from such direct and immersed exposure, would indirect exposure through the media perhaps have a more persistent bearing on the results? To what extent would TV influence lectal awareness? Let us hypothesize that the more TV the children watch, the more exposure to lectal and social variables (and more exposure > more awareness > better identification). To test this possibility we correlated the self-esteemed number of hours the 12-13 year-olds watch TV/week with the results obtained in the tests. Table 28 shows the results for experiment 1 at the level of autonomous community.

   As the table shows, the fewer hours of TV/week the children watch, the better the results are, and the more they spend, the worse the result become. By way of example, none of the children who claimed that they hardly

watch TV or do not watch TV excessively (below 2 hours or from 3-7 hours per week) obtained no correct results in the test. On the contrary, these youngsters obtained the best results in the test: while one of the children who claimed to spend more than 15 hours a week in front of the telly obtained 5 correct hits out of 8 possibilities, so did 6 of the less addicted. While it cannot be ruled out that watching TV was regarded by the children as a negative aspect when evaluated by an adult and that their self-evaluation was influenced by this factor, on the basis of the data at our disposal we must conclude that to the extent that exposure from the media is a relevant factor, at least the quantitative dimension is less important than hypothesized.

But there is more data to be discussed in relation to social awareness. The 150 children were also asked the following question for all fifteen speech samples: *What other people or characters do you know who speak this way?* The question was included in the response form in order to elicit correspondences between the speech fragments and social categorizations, just as the questions regarding travelling and TV we just discussed did, just that in a more indirect fashion. Which type of people or characters did the children know who spoke like the speakers in the fragments?

Qualitatively speaking, the 12-13 year-olds had the following responses for Andalucía:

· our English teacher Maria Jose (11 out of 50 listeners)
· relatives, neighbours, friends
· María del Monte, Isabel Pantoja (Andalusian singers)
· Sergio Ramos (football placer from Real Madrid)
· Cristina García Ramos (TV presenter)

For the American speaker, the replies were as follows:

· relatives in Florida
· a summer camp
· a character in Los Serrano (a TV series)
· David Beckham (British, not American)
· Hilary Duff

For the Argentinean speaker the responses included:

12-13 year-olds
· my dentist
· a friend
· Leo Messi (football player in F.C. Barcelona, 4 responses)

· Maradona and other Argentinean players (3 replies)
· characters in the TV series Hospital Central and Rebelde (9 responses)
8-9 year-olds
· Mateo (2 responses)
· Mateo´s parents (6)
· characters in the TV series Rebelde (1)

What do these examples tell us about social awareness in relation to lectal categorization? The children do seem to build up a lot of knowledge from their most immediate surroundings, i.e. the school and family environment and the TV. For the latter factor it seems to matter most *what* they watch and not *how much* they watch. In early 2007 when the tests were distributed, the Argentinean youth series "Rebelde Way" (together with its Mexican follow-up sister series "Rebelde") was very popular amongst Spanish teenagers. The responses of the 12-13 year-olds reflect this fact very well against the single reference to the series on the part of a listener from the 8-9 age group. In turn, 8 listeners aged 8-9 mentioned "Mateo" or "Mateo's parents" as an example of someone speaking with an Argentinean accent. Mateo was a child with an Argentinean background who happened to be in one of the classes in which the test was placed. Popular singers and football players whose voices are heard in the media also rated high in the response form. The media do play a role, then, as do personal experiences with people in the most immediate environment. The comparatively bad results for the non-native accents can in part be explained by the fact that foreign languages are invariably dubbed on TV in Spain. An additional contextual factor involves L2 teaching. Bilingual educational programmes in which native English teachers participate as auxiliaries are becoming increasingly widespread in Spain and English speech is thus heard by the children from the time when most of them begin school, i.e. at the "pre-school" or kindergarten level at the age of 3. However, most of the children who participated in the experiments in 2007 had begun foreign language instruction in English at the age of 6, in their first year of compulsory education – and more often than not their English teachers did not speak with an English accent.

In Table 29 we specify the quantitative references made by the children to the accents identified (excluding the Spanish reference accents). It is obvious that these results do reflect the success rate in a quite transparent manner. Regardless of objective distinctive and exclusive features and linguistic frequencies, the children rated the accents in accordance with schemas built up around familiar people and characters in their immediate expe-

riential world. The knowledge they rely on stems from personal experience and from the media alike, but in any case it is experientially grounded and based on social salience rather than on inherent linguistic characteristics.

Table 29. Familiarity as a subjective distance

	Distinctive features	Exclusive features	Frequent features	Linguistic awareness	Social awareness
Andalucía	4	1	2	3	**45**
Galicia	3	1	2	2	**8**
Canarias	4	1	2	3	**1**
Argentina	3	1	1	3	**41**
Mexico	2	1	1	2	**20**
France	4	1	3	4	**10**
Germany	3	0	3	3	**6**
Br. Eng.	4	1	3	5	**5**
Am. Eng.	3	1	3	2	**5**

To end this section let us address the following question: would lectal identification then at least be structurally determined in the sense that lectal disambiguation is mediated by constructs larger than the phoneme, such as lexical items? I.e. we realize that phonetic variants pertain to a given lect because the words in which they occur constitute a frame (or model or construction) that determines the feature as a (socially relevant) variant of a given phonemic category? In the L1 experiment, when at the very beginning of the text even the 6-7 year-olds heard the Spanish word <oye> (`listen´) pronounced with intervocalic [ʃ] or [ʒ], they immediately made use of their pencils to categorize the accent in question as Argentinean. Was this because the lexical context of the word <oye> helped them realize that the consonant corresponded to templates of a phoneme with socially derived variants, out of which [ʃ] and [ʒ] are typical instantiations or Argentinean lects?

In a paper entitled "Phonological development: toward a ''radical'' templatic phonology" Vihman and Croft (2007) extend the theoretical framework of radical construction grammar to phonology. Vihman and Croft argue that the word is the basic unit of phonological representation just as constructions are basic and syntactic categories of particular units are derived from the constructions. In other words, phonological disambiguation is lexically mediated:

Labov's research on a single individual's productions of vowel tokens (Labov 1994, inter alia) demonstrates that individual exemplars of one phoneme will be included in the phonetic range of another phoneme: for example, some exemplars of /æ/ will occur in the range of exemplars of /ɛ/. How does a speaker know that those tokens are exemplars of /æ/ and not /ɛ/? This question cannot be answered in a purely segment-based approach to phonological representation. If one begins with segments, one must have a definition of those segments that is either ultimately phonetic, or else purely arbitrary (i.e., a particular exemplar is stipulated to be an exemplar of /æ/ even if its actual realization is [ɛ] in purely phonetic terms). On the other hand, if one begins with words as phonological units, then the question can be answered and the paradox is solved. The phonetically outlying token is an exemplar of /æ/ because it is part of a specific word, and other occurrences of that word contain exemplars that cluster around the central phonetic tendency for /æ/. How is the word identified as the same word? The word is of course identified as the same by its meaning in the context of use, linked to prior occurrences of the word with that meaning in similar contexts of use.

Vihman and Croft thus argue in favor of an approach to phonological prototype categories (cf. *exemplars that cluster around the central phonetic tendency of /æ/*) according to which phonological classification is determined by units, or constructions, larger than its components. In the case of lects, however, I would like to argue that the lexicon *facilitates* disambiguation, but fails to *determine* it. To illustrate the difference, imagine the following situation:

On a university campus you catch a crowded bus and notice a group of students engaging in lively conversation at the rear end. You cannot really hear what they say because of the background noise, but you capture sounds and tones that tentatively identify the group of students as speakers of, say, Icelandic or Dutch (or any other language you are familiar with). It is only when you reach the rear end of the bus to get out at your destination that you realize that the students in question are not speaking in their mother tongue but in the native language of the country you are in (say, England or Spain). Like the children in experiment 2 you were able to identify the L1 language of the speakers, but in order to do so you did not rely on structural constructions at the level of words or clauses. The link went directly from a stored linguistic stereotype to the unclassified speech tokens that you heard.

If we can agree that this is not a far-fetched example, but rather a situation that many of us have actually experienced, we draw the tentative con-

clusion that while phonological disambiguation might well be lexically mediated, lectal disambiguation seems to follow a different path. Here the link is likely to go directly from a stretch of unclassified speech to a linguistic stereotype stored in a shape which reflects the experiential stage reached by the identifier.

5.   Theoretical implications

In this chapter we have charted emergent patterns of lectal acquisition in children aged 6-13 and witnessed the extent to which identification becomes increasingly more accurate and how type/token correspondences remain constant across the three age groups. We furthermore saw the extent to which the children paid attention to phonetic and suprasegmental detail and were able to extrapolate from one instantiation (or token) to the (linguistic stereo-) type and from this to members of the socially related categories.   We conclude that structured models of lectal variation gradually emerge in the course of the child´s first six years and evolve so as to become both qualitatively and quantitatively more refined and effective when the child reaches pre-adolescence.

In section 1 we predicted that in linguistic stereotype formation structure would be determined by usage and that purely formal characteristics (such as phonetic salience and contrast) would have fewer effects on correct identification than relative social salience (such as social stereotyping). The data confirms this prediction: linguistic stereotypes build on patterns of actual usage in the child's experiential world rather than on intrinsic linguistic features. The central images of lectal categories that enable language users to correctly and effectively identify a stretch of speech as an instance of "accent X" seem to emerge when the child processes and associates a stored amount of specific data relating to speech styles and social events, accumulated in everyday experience.

In this chapter we have thus argued that lectal identification works fast and effectively because lectal schemata that relate to the social categorizations they derive from are gradually built up on an experiential basis. In this respect we have concluded that successful identification does not depend primarily on the existence of especially salient structural features in a given variety, but rather on unique combinations of linguistic features associated with socially salient and stereotyped categorizations.

References

Bybee, Joan
  1999   Usage-based phonology. In *Functionalism and Formalism in Linguistics, Vol. I: General Papers*, Michael Darnell, Edith Moravcsik, Fredrick Newmeyer, Michael Noonan, and Kathleen Wheatley (eds.), 211-242. Amsterdam/Philadelphia: Benjamins.
  2001   *Phonology and Language Use*. Cambridge: Cambridge University Press.
  2003   Mechanisms of change in grammaticization: The role of frequency. In *The Handbook of Historical Linguistics*, Brian D. Joseph, and Richard D. Janda (eds.), 602-623. Oxford: Blackwell.

Kristiansen, Gitte
  2001   Social and linguistic stereotyping: a cognitive approach to accents. *Estudios Ingleses de la Universidad Complutense* 9: 129-145.
  2003   How to do things with allophones: linguistic stereotypes as cognitive reference points in social cognition. In *Cognitive Models in Language and Thought*, CLR 24. René Dirven, Roslyn M. Frank, and Martin Pütz (eds.), 69-120. Berlin/New York: Mouton de Gruyter.
  2006   Towards a usage-based cognitive phonology. *International Journal of English Studies* 6 (2): 107-140.
  2008a   Style-shifting and shifting styles: a socio-cognitive approach to lectal variation. In *Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems*, CLR 39, Gitte Kristiansen, and René Dirven (eds.), 45-88. Berlin/New York: Mouton de Gruyter.
  2008b   The group as a variable in the development of cognitive schemata. In *Body, Language and Mind, Vol. II*, CLR 35.2, Roslyn M. Frank, René Dirven, Tom Ziemke, and Enrique Bernárdez (eds.), 409-432. Berlin/New York: Mouton de Gruyter.

Nathan, Geoffrey S.
  1996   Steps towards a cognitive phonology. In *Natural Phonology: The State of the Art*, Bernhard Hurch, and Richard Rhodes (eds.), 107-120. Berlin/New York: Mouton de Gruyter.
  2006   Is the phoneme usage-based? – Some issues. *International Journal of English Studies* 6 (2): 173-194.

Nunberg, Geoffrey D.
  1978   *The Pragmatics of Reference*. Bloomington, IN: Indiana University Linguistics Club.

Pierrehumbert, Jeanette
  2001   Exemplar dynamics: word frequency, lenition, and contrast. In *Frequency Effects and the Emergence of Linguistic Structure*, Joan By-

bee, and Paul Hopper (eds.), 137-157. Amsterdam/Philadelphia: Benjamins.

Purnell, Thomas, William J. Idsardi, and John Baugh
    1999   Perceptual and phonetic experiments on American English dialect identification. *Journal of Language and Social Psychology* 18: 10-30.

Taylor, John R.
    1990   Schemas, prototypes, and models. In *Meanings and Prototypes: Studies in Linguistic Categorization*, Savas L. Tsohatzidis (ed.), 521-534. London: Routledge.
    1995   *Linguistic Categorization. Prototypes in Linguistic Theory*. Oxford: Clarendon Press.
    2002   *Cognitive Grammar*. Oxford: Oxford University Press.

Van Bezooijen, Renée, and Charlotte Gooskens
    1999   Identification of language varieties. The contribution of different linguistic levels. *Journal of Language and Social Psychology* 18 (1): 31-48.

Vihman, Marilyn, and William Croft
    2007   Phonological development: Towards a "radical" templatic phonology. *Linguistics* 45 (4): 683-725.

Investigations into the folk's mental models of linguistic varieties[1]

Raphael Berthele

Abstract

Most sociolinguists today deny that there is an intrinsic link between the formal properties of languages and dialects and non-linguists' language attitudes. In this study, the question of an intrinsic contribution of vowel phonology is asked, based on data collected in an exploratory experimental picture mapping task. The task is used to investigate laypeople's mental models of languages and administered to speakers of Swiss German dialects. A particular focus lies on the participants' perception of dialectal variation within the German varieties. Additional interview data on the participants' perception of dialects is discussed and the quite consistent evidence for a link between closed and high vowels and angular, spiky visual forms is discussed based on the research results on synaesthesia. The chapter also discusses cultural models of standard and non-standard languages and integrates the perceptual and cultural aspects into a holistic perspective that aims to deepen our understanding of mental and cultural models of linguistic varieties and sociolinguistic prestige.

Keywords: folk linguistics, synaesthesia, perceptual dialectology, sociolinguistics, attitudes, cultural models, metaphors

1. Introduction

1.1. Sociolinguistics and Cognitive Linguistics

Sociolinguists have developed very impressive methods of describing patterns of lectal variation both from a qualitative and a quantitative perspective. Whereas the initial agenda of sociolinguistics was to describe correlations between social strata and linguistic forms or "codes", a specific

branch of sociolinguistics has more recently been focusing for a number of years on a more fine-grained dynamics, such as the convergence and divergence of language choices in social networks (Milroy 1992), as part of processes that can be understood as "acts of identity" (cf. le Page and Tabouret-Keller 1985). However, the attitudinal values attributed to languages, varieties and styles remain very much related to the vertical dimension as expressed in the Fergusonian (1959, 1991) distinction between High- and Low-variety or the concept of prestige (Labov 1966, Trudgill 1972). Despite the numerous investigations into sociolinguistic attitudes and evaluations of standard and non-standard speech, sociolinguists are surprisingly uninventive in explaining the puzzling fact that speakers continue using certain varieties despite their obvious and openly admitted lack of prestige (cf. Niedzielski and Preston 1999: 101). On the whole, sociolinguistic theorizing about the attitudinal component of the sociolinguistic dynamics has been relatively poor. Labov's (1966) notion of "covert prestige" is probably the best known attempt to account for the attitudinal basis of the maintenance of non-standard speech, and this concept will be further discussed in section 2.2 below.

In this chapter, I argue that concepts from cognitive linguistics (CL) are useful for a better understanding of folk models of linguistic variation, standardization, and sociolinguistic values of varieties. This can be seen as an attempt to reply to Preston's point that we need a "folk theory of language":

> [W]hat underlying beliefs, presuppositions, stereotypes, and the like lie behind and support the existence of language attitudes? Ultimately, it seems to me, this will require us to give something like an account of a folk theory of language […]. (Preston 2004: 41)

Since one of CL's most remarkable achievements is to offer theories of (natural) categorization and cognitive models (ICMs, metaphors, and the like) it seems sensible to use these theories also in the realm of folk models of language. More specifically, in my view, CL should be particularly helpful for the study of the cognitive component of language attitudes (cf. Lasagabaster 2005).

In this contribution, as a small step in this direction, a "cognitively inspired" way of eliciting gestalt-based mental models of languages/varieties is demonstrated, and consistent patterns of attribution between visual traits and phonological features are presented. Moreover, I suggest that some of the evidence is consistent with what is generally called the (very controver-

sial) "inherent value hypothesis" (Giles, Bourhis, and Davies 1975), i.e. the claim that some language attitudes are due to inherent features of the systems and not only due to culturally or socially imposed norms, stereotypes or socio-cultural connotations. It is important to note that this focus on inherent and thus potentially universal features does not automatically entail a universalist, non-relativist position (as the one that seems to be taken by Lakoff 1987: 268; cf. the discussion in Kristiansen 2008: 410): Rather, as other authors have emphasized before, the goal here is to strive for an integration of embodiment, perception and cultural and social experience within a holistic construal of linguistic competence and usage. In other words, one of the goals of this chapter is to show how potentially universal perceptual mappings of sounds and forms interact with cultural and other mental models of social or ethnic groups, languages and varieties. The seemingly opposite positions of inherent values vs. cultural stereotypes regarding language attitudes in my view is a wrong opposition, since both domains contribute their share to attitudes and mental models of language in a conspirative manner, as will be argued below.

1.2. Cognitive Models, Cultural Models, and other Theoretical Assumptions

In his paper on cultural models of language standardization, Geeraerts (2003) discusses two fundamentally different perspectives on language standardization. The rational stance on language and language standardization construes language primarily as a means of communication. The standard language is best suited for this function for several reasons: it is (supposedly) geographically and socially neutral and sufficiently general, it can be used for the verbalization of any topic, and it is connotationally neutral and thus in sum the ideal means of social emancipation in the pre-modern and modern societies. Variation and non-standardness, on the other hand, is an impediment to emancipation. A prototypical instantiation of this model, also discussed by Geeraerts, is the exclusive enforcement of Standard French in revolutionary and post-revolutionary France. The Romantic stance can be related not only to the romantic view of "indigenous" and "authentic" local languages, but also to more modern sociological views (e.g. Bourdieu 1982) on the varieties of dominant social strata. From this perspective, the standard language is construed as a means of exclusion and

oppression; the dialects are the "real" and authentic expression of local culture and individual and local identity.

In Berthele (2008, 2001) I used the theory of conceptual metaphors (Lakoff and Johnson 1980) in order to get a better grip on language policy debates in different countries. These debates typically address issues such as official or national languages, the choice of the main language(s) of instruction or more generally the order and choice of foreign language instruction in schooling. The metaphors I identified in these language policy debates can be related to Geeraerts' two models, as shown in Table 1.

Table 1. Metaphors used in language policy debates (cf. Berthele 2008, 2001) and cultural models of standardization (Geeraerts 2003)

Romantic Model – Language Is …	Rationalist Model – Language Is…
- the Soul of A People	- a Tool, Key, Bridge
- a Bond, Glue, Tie	- a Bond, Glue, Tie
(nation, speech community)	(nation, humanity)
- Home	- Art, a Balanced Structure, a Building
- a Resource (cultural, heritage)	- a Resource, raw material
- a Barrier: Standard blocks the	- a Barrier: Dialects/Minority Ls as means
vertical mobility of L-Minorities	of (self-)exclusion

As Table 1 clearly shows, some metaphors can be used in both models, e.g. the very powerful idea of LANGUAGE IS A BOND can be applied with different levels of granularity, either to a national society as a whole in the rationalist perspective or to the local speech community in the romantic perspective. As Geeraerts (2003: 28) points out, both models can converge in nationalistic ideologies, with a romantic flavor for the ethnically based "Volksnationalismus" and a rationalist flavor in the case of the "Staatsnationalismus". Correspondingly, the Bond-metaphor applies both to ethnic groups that are equated with nations and to national groups that can be multi-ethnically composed. In similar ways, the mapping of LANGUAGE IS A BARRIER can be applied both to the standard and to non-standard varieties, depending on the general stance being taken.

Other metaphors are clearly related to one of the two models, such as the idea of a language being a perfectly designed structure such as, e.g., an architectural masterpiece (building), a piece of art or – more abstractly – a perfectly logical theory. Again, particularly the French metalinguistic dis-

course gives perfect examples of this construal, e.g. in the very widespread idea of French being THE logical and clear language par excellence ("Ce qui n'est pas clair n'est pas français", attributed to Antoine de Rivarol; cf. also Swiggers 1990).

Whereas the two models outlined by Geeraerts, both essentially cultural (and more specifically European) in nature, clearly belong to the realm of "language ideologies", some of the metaphorical mappings listed in Table 1 also draw on the experiential domain. E.g., the construal of LANGUAGE AS A BARRIER clearly relates to the early experience of any language learner (first, second or nth language) that the lack of proficiency often entails no or only partial achievement of communicative goals and thus imposes serious constraints on the fulfillment of basic needs. Moreover, the construal of LANGUAGE AS A TIE OR BOND clearly relates to the again very early experience of language as one of the most important domains of cooperative social practice. If anything were to materially substantiate the intrinsically abstract nature of a social group in the first place, then it would be converging cooperative social practice such as the cooperative use of symbols in social interaction.

In the remainder of this contribution, I will discuss data collected in a folk linguistic study (cf. Berthele 2006) that relates directly to the models outlined here. Bloomfield reportedly referred to non-linguists' evaluations and ideas about language by using the depreciative term "stankos". One of my goals will be to demonstrate that the folk's descriptions of varieties and languages expose not only more 'descriptive accuracy' than linguists generally assume, but also that surprisingly consistent perceptual patterns regarding e.g. the relation between the phonology of varieties and the evaluative-attitudinal comments can be discovered. These patterns will be shown to lead to the hypothesis that the mental models of languages and varieties are not always purely ideological, but that there can be an important experiential component that feeds into the construction and social reproduction process of these models.

2.   The Study

The data discussed below were collected in the Swiss-German context. The particular advantage of the Swiss-German language situation is that it is characterized by a relatively high degree of ethnolinguistic vitality of local dialects – at least if compared to the surrounding Romance or Germanic

areas. Swiss-Germans are thus more or less bilingual in two closely related systems, the respective base dialect and the Swiss variant of the German standard language, and the relatively clear functional distribution of usage of the two systems is commonly associated with the concept of diglossia (Ferguson 1959, 1991). What is particularly important for the present study is that – at least on the phonological and morphological level – there is no such thing as a Swiss German Koiné, i.e. a leveled out form of an Alemannic dialect (cf. Christen 1998). Therefore, we still find a considerable amount of phonological variation across the Swiss German speaking area, and Swiss Germans appreciate this diversity, are generally very aware of it, and love to mock (deprecatingly, but sometimes also admiringly) their compatriots' dialects.

In 2004, we[2] carried out standardized data elicitation sessions with 45 adult participants, all of them native Swiss Germans between 19 and 87 years old. These participants were asked to participate in 5 tasks:

1. Map elicitation task: Participants had to draw by hand dialect/language maps of Switzerland
2. Dialect imitation task: Participants were asked to imitate dialects they know
3. Dialect recognition task: Participants were played dialect samples and they had to recognize the region of origin of the speaker
4. Visual stimulus mapping task ("Bubble Task"): Participants were presented a series of hand-drawn bubbles (see Figure 1) and asked to attribute varieties and/or languages to these bubbles
5. Attribution task: While doing task 4, participants were asked to characterize the languages/varieties they mapped onto these bubbles

2.1. Bubble task and attribution task

The tasks that are particularly interesting for the issues raised in section 1 are tasks 4 and 5. Our participants were first handed a sheet of paper showing the bubbles in Figure 1. Then, the participants were requested to do two things:
1) Can you assign any varieties/dialects/ languages to any of these bubbles?
2) Can you describe the varieties/ dialects/languages you are referring to?
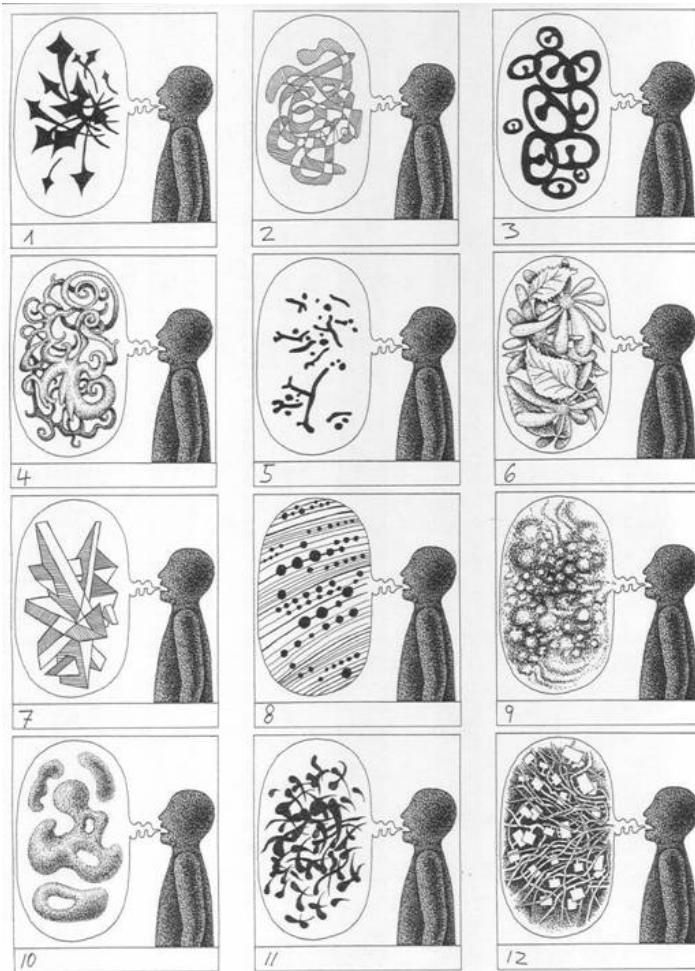
Figure 1. The bubble task. Artwork by Andreas Gerber, Bern

In order to facilitate the task, the researcher gave the participant a form with prepared columns for the number of the bubble, space for the labels of the languages/varieties, and some space for the description of each entry. Although this mapping of visual form onto languages is – from an objectivist and matter-of-factly point of view – quite nonsensical, almost everybody spontaneously and joyfully started to suggest such associations (3 Participants out of the 45 did not participate in this task - only one of them categorically refused to do the task because he felt it was nonsensical). The

study was explorative in the sense that I was not sure whether there was such a thing as a perceptual mental "gestalt" for languages and varieties that speakers know. The spontaneous reaction of our participants seems to suggest that this idea might not be totally wrong.

2.2. Attributions, characterizations, and cultural models

As we will see below (e.g. in Table 3), there are various languages and varieties that are attributed to the bubbles, and since the design does not force choices out of the participant but lets them associate freely, the data of the different participants are not comparable in a straightforward way.

The first analysis to be presented here zooms in on the two most frequently mapped varieties in the data. Out of the 45 participants, 35 attributed Standard German to at least one of the bubbles and 37 participants attributed the Alemannic dialect of Bern to at least one of the bubbles. With regard to the two models laid out above, it is thus interesting to take a closer look at three aspects:

1) Which are the typical bubbles chosen by the participants for these two varieties?
2) Which are the typical attributes given to characterize the two varieties?
3) Can these mappings be related in any way to the cultural models of language standardization?

Table 1 lists the 3 most frequently mapped bubbles (numbers refer to the numbers in fig. 1) per variety as well as all attributes that have been given more than once for the particular variety. The most striking difference between the two mapping patterns is that the Bern variety is consistently associated with round, organic, or floral forms (the three picture stimuli listed top left in Table 2 make out about 70% of all the mappings for this particular variety), whereas the standard language is preferentially associated with angular and pointed forms in more than 70% of all mappings. Most attitude studies in German-speaking Switzerland have shown that there are negative attitudes towards the standard language (Häcki Buhofer and Burger 1998), a result which is unsurprising to people familiar with the local landscape. This widespread negative attitude towards the standard language is in sharp contrast with many other European countries (e.g. the Netherlands; cf. van Bezooijen 1997).

Table 2. The three most frequently mapped bubbles and the verbal attributes (N>1) for the two most frequently described varieties. Figures in brackets represent absolute numbers of occurrences

Berndeutsch (Bern dialect)	Standard High German
Bubbles:	Bubbles:
10 (14, 36%)	7 (15, 43%)
6 (9, 23%)	1 (6, 17%)
3 (4, 10%)	12 (4, 12%)
soft (7)	clear (10)
bloomy (5)	fast (5)
slow (5)	jagged/chiseled/angled (6)
broad (4)	complex (4)
homey, homelike (3)	sharp, pointy (3)
calm (3)	verbose (3)
self-contained (2)	rule-governed/regular (5)
warm-hearted (2)	minute (3)
clear (2)	impersonal (2)
big, large (2)	real language (2)
round (2)	cragged (2)
beautiful (2)	

The attributes listed in Table 2 now allow a more fine-grained analysis of the mental models that seem to be related to the two varieties: The Bern dialect on the one hand seems to be construed as a cozy, warm, and esthetically pleasant variety. The standard, on the other hand, is not only sharp and chiseled, but also much more clearly structured and more than once explicitly labeled "a real language". In the visual domain, there is a rather systematic correspondence to the European topos of the 'good' standard language being rule-governed, logical, and optimal for communication given its precise nature: participants clearly tend to chose clear and precisely drawn, sharp figures. Thus, we can confirm the finding from other studies (e.g. Niedzielski and Preston 1999: 22) that the standard is appraised as a well-structured and rule-governed system, whereas the dialect does not have or does not need any such clear norms, is rather unstructured, spontaneous and somehow "authentic" in its uncultivated state. It thus seems that we can indeed find reflections of the romantic and the rationalist models.

From the romantic point of view, the dialect is the expressive, emotional variety (*bloomy*, *beautiful*), it represents intimacy and familiarity (*homey*, *calm*, *warm-hearted*), and it relates to the geographic and social space of proximity with a strongly expressed local identity (*self-contained*). Dialects, by definition languages with strong local cultural roots, are acquired and not learnt – in opposition to the Standard whose system has to be constructed through hard work, e.g. in school – but passed on as a heritage or *home*. The standard, on the other hand, from a romantic point of view, is artificial, *impersonal* (cf. Table 2) and optimized for speed (*fast*).

As already discussed above, the rationalist view on the standard language corresponds perfectly well with many attributes found in our data, such as the uniformity/standardness of the *rule-governed*, *real language* and its *regularity*. Other attributes that match the model proposed by Geeraerts perfectly well are those which convey the notions of effectiveness and speed (*clear, fast, verbose*) and the domain-generality (*non-personal*). Thus, the rationalist stance is represented in our data by a number of attributes that characterize the standard language as a regular, complex, efficient and logical construction as opposed to the "random" or "undomesticated" dialect. The relative consistency of these attributions can also be highlighted by the fact that hardly any property is attributed to both varieties by our participants: Only 4 types out of a total of 68 for these two varieties appear for both the Bern dialect and the standard language (*clear, angled, chiseled, clumsy*).

The overall picture that emerges from the analysis of the mapping and attribution task can be summarized as follows: There is support for both of the models that Geeraerts (2003) proposed. The two most often described varieties, one a standard language, one the Alemannic dialect of the capital of Switzerland, clearly tie in with the Rationalist and the Romantic view of standardization. The data make it very clear that, just as Geeraerts (2003) has pointed out, participants do not choose either to apply the Romantic or the Rationalist model to all languages or to language in general. What we find is that the two models – together – form a complex model in the sense of Lakoff's cluster models (cf. Lakoff 1987: 74). The cluster model contains a rationalist and a romantic sub-model (cf. fig. 2). Both of those sub-models selectively apply to the prototypical dialect and to the prototypical standard language.

It is important to note that, at least in my view, these models are by no means static and brassbound ways of construing languages and varieties. They have to be understood as dynamic entities of situated cognition, pro-

viding visual and other metaphorical source domains for the construal of esthetic judgments and argumentative trains of thoughts about language. It is neither the case that individuals have to select categorically one of the two models as a guiding model, nor do individuals always have to apply the same models in a similarly balanced way. However, particular cultural discourse contexts, such as the "republican" ideology in France, can clearly favor one model at the expense of the other, and there is no doubt that such entrenched cultural ways of construing language leave their traces in the minds of the members of the speech community.

competing/partial models of a "cultural cluster model":
-> Romantic component ~ dialects
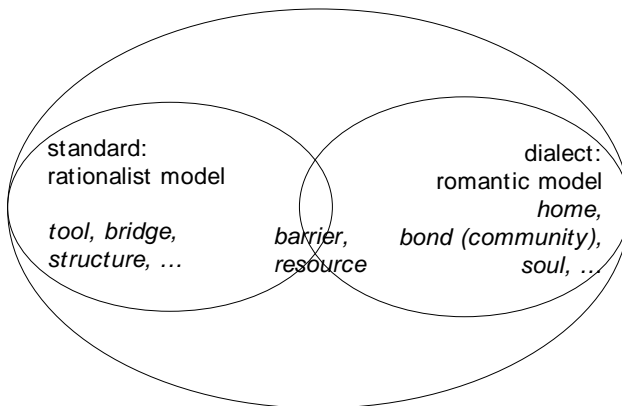-> Rationalist component ~ standard

Figure 2. The cluster model of languages and dialects; integrating the cultural models of standardization and some metaphorical models

The concept of this cluster model also accounts for the often perceived internal inconsistencies in the folk's ideas about language – or for the general finding that people's mental models of all kinds of objects are inconsistent (cf. the pastiche model, Collins and Gentner 1987). In our case, the analysis presented above can easily account for a fact that has only been poorly understood in sociolinguistics: if the hypothesis of the cluster model is correct, then it comes as no surprise that people can attribute prestige both to the standard language and to the dialect. The construct of "covert prestige" (Labov 1966, Trudgill 1983) is the sociolinguistic concept that

comes closest to the multi-dimensional construal emerging from our data. However, the prestige of the sociolect or dialect is probably not necessarily covert; its overtness or covertness simply depends on the context of elicitation and is thus a question of perspective. Obviously, the Norwich working class males (cf. Trudgill 1983: 177) did not, in the presence of the researcher, overtly admit that they actually have a positive attitude towards their native variety. But this does not automatically allow for generalizations about the overtness or covertness of sociolinguistic prestige. There is no doubt that the Swiss German situation is significantly different from the British context. But there is no reason to assume that other data elicitation techniques would actually reveal openly positive judgments about non-standard varieties also in contexts where the non-standard might have a harder life than in German-speaking Switzerland. As shown above, approaches from cognitive linguistics and their application to the data collected in this project give a much more fine-grained picture of linguistic value judgments and language loyalties than the traditional vertical axis applied in sociolinguistics (high vs. low; prestigious vs. non-prestigious; standard vs. non-standard). The two varieties discussed above obviously both bear sociolinguistic prestige, but each one of a different kind: The dialect is the variety with the "warm" prestige of the language of proximity, it carries solidarity and attractiveness, whereas the standard is the high-status bearer of the "cold" prestige of the "real", logical and well-designed artifact. This reflects quite nicely the result of Hogg, Joyce and Abrams (1984) where Swiss German is tied to the dimension of solidarity whereas the standard language is related to status. Whether a particular manifestation of prestige is overt or covert is not an intrinsic quality of linguistic prestige but rather a question of the perspective of the observer with respect to the observed and to the reference points that are activated.

2.3. Attributions and phonological inventories of varieties

In the preceding section I have compared the mappings regarding the two (statistically) most important varieties in our sample. Implicitly, the Bern dialect has been treated in a metonymical fashion as representing THE dialect par excellence. This has been justified by the relatively open nature of the task and the constraint to work only with varieties/languages that have been frequently characterized by the participants. However, there are – as ever – enormous differences when it comes to the esthetic, ethnic and cul-

tural evaluation of different (Swiss German) dialects. Instead of adopting the simple binary opposition of dialect and standard it is important to take into account that there are various more or less stereotypical models of many dialects (cf. Ris 1992; Werlen 1985). These stereotypical evaluations of the different dialects have been described extensively, and the general point of view taken by the researchers is the – rather uncontroversial - hypothesis that negative or positive attitudes towards certain varieties are due to cultural stereotypes (connotations) and imposed norms (cf. Giles, Bourhis, and Davies 1975) and not to inherent characteristics of the languages/varieties.

Table 3. Frequencies of mappings per stimulus, relative to the total number of respondents (N=42). Only varieties which were attributed by 10 or more participants are considered in the analysis

stimulus #	French	Berne dialect	Zurich dialect	Italian	Basel dialect	Standard High German	Wallis dialect	St. Gallen dialect	English	Bavarian dialect	Uri dialect	Aargau dialect	Swiss German	Suabian dialect
1	0%	2%	21%	10%	12%	14%	2%	24%	0%	5%	0%	5%	0%	2%
2	5%	0%	2%	12%	5%	5%	5%	5%	10%	5%	7%	2%	0%	0%
3	10%	10%	5%	5%	14%	2%	5%	5%	14%	10%	2%	5%	2%	0%
4	24%	7%	7%	7%	14%	2%	2%	0%	5%	7%	0%	0%	5%	10%
5	5%	2%	5%	2%	5%	0%	0%	5%	10%	0%	12%	5%	0%	10%
6	17%	21%	2%	10%	5%	2%	21%	0%	2%	12%	2%	0%	2%	5%
7	5%	5%	21%	2%	5%	36%	5%	19%	0%	0%	10%	0%	7%	0%
8	7%	2%	2%	10%	2%	10%	5%	2%	7%	0%	0%	10%	5%	0%
9	12%	7%	2%	5%	7%	2%	7%	5%	2%	2%	2%	10%	12%	2%
10	2%	33%	0%	0%	5%	0%	5%	2%	5%	7%	10%	2%	2%	2%
11	2%	2%	5%	17%	5%	0%	14%	2%	2%	5%	0%	2%	2%	2%
12	7%	0%	17%	5%	5%	10%	7%	7%	7%	2%	0%	2%	5%	2%

In this section, however, I will at least partially make an argument for the possibility of an inherent linguistic component of our mental models of language (cf. for a similar line of thought Cuonz 2008). This will be done by comparing the different mapping patterns for two major Swiss German

dialects, two dialects that are both geographically and linguistically op-
posed to each other.

    Before discussing these details, it is useful to look at the overall "land-
scape of mappings" in our sample. In order to get a general overview of the
different languages and varieties that were mapped onto the visual stimuli
in Figure 1, we can take a look at relative frequencies and relative distances
between the varieties that can be calculated via statistical scaling proce-
dures. First, Table 3 represents those varieties that were mapped at least 10
times to any one of the stimuli depicted in Figure 1. Table 3 gives the read-
er an idea of the languages and varieties that have been associated with the
stimuli by the participants. The level of language-categorization is differ-
ent, we find rather fine-grained lectal categories such as "Uri dialect", "Ba-
varian dialect" next to the quite unspecific global term "Swiss German"
which does not refer to a specific variety but rather to all the different
Swiss German varieties, and obviously there are also labels for "Abstand"-
languages such as "English" and "French".

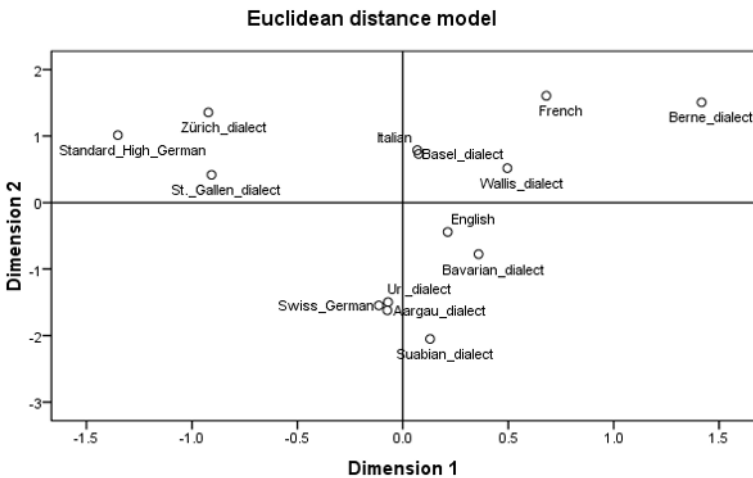*Figure 3.* MDS analysis of the data given in Table 3

The values given in Table 3 can be transformed into spatial distances and
coordinates by applying the multidimensional scaling procedure (MDS;
Kruskal and Wish 1991). The 2-dimensional output of this can be seen in
Figure 3. As usual, the dimensions in MDS output are not meaningful as
such, they simply represent a measure of relative proximity with respect to

the selected variables. In our case, this chart shows to what degree the varieties/languages in Table 3 are mapped onto similar visual stimuli in the bubble task. Thus, data points which are close in Figure 3 are varieties that were characterized largely by the same visual stimuli. Data points that are far away from each other represent varieties that were hardly ever mapped onto the same visual stimuli.

It is not possible to interpret the distances between all the points in Figure 3 in a straightforward manner. But the figure can be helpful for the analysis of particular pairs or groups of varieties. For instance, it is quite obvious that the two varieties discussed in the preceding section, Bern Swiss German and Standard High German, are located at opposite ends of dimension 1. This backs up the conclusion already drawn above that the gestalt representation of these two varieties could hardly be more different. We will come back to a possible meaningful interpretation of dimension 1 when discussing the different mappings of Bern dialect and St. Gallen dialect below. For the present discussion, however, it is interesting to notice that the plot in Figure 3 shows one relatively well delimitated cluster in the upper left quadrant. In this cluster we find eastern Swiss German dialects (Zürich and St. Gallen) as well as the German Standard language. As the discussion below will show, all of them are systematically associated with spiky, chiseled, angular forms. In the lower middle part of the plot there is cluster of varieties that is less well-defined and not so easily interpretable: here we find the generic roofing term for all Alemannic dialects spoken in Switzerland (*Swiss German*) as well as a very un-prototypical variety of these dialects: Aargau dialect. This category is generally perceived as a cluster of quite different local dialects in the transition zone between the urban centers of Bern, Zürich and Basel (cf. Siebenhaar 2000) that is carved up by some of the most important isoglosses that categorize the Alemannic dialects. Therefore, neither in the dialectological literature nor in folk perception do we expect there to be such a thing as a focused representation of the Aargau dialect. It comes thus as no surprise that the bubble choices associated with the *Swiss German* and the *Aargau* category are distributed across several bubbles, with no stimulus reaching more than 12% of the vote of our participants. This is in sharp opposition to the choices for Bern or St. Gallen dialect. Here, at least one stimulus gets 24% or even 33% of the choices. This cluster at the lower middle of the plot is not easily definable, as is the case for dimension 2 in general, but there is a tendency to find not prototypical (Swiss German, Aargau) and geographically or genetically more distant (Suabian, Bavarian, English) and quite

exotic alpine (Uri) languages/dialects clustering together. What characterizes the dialect categories in this cluster could be their (maybe with the exception of Uri) status of non-prototypical Swiss German dialects. The upper part of this middle zone, on the other hand, unites two Romance languages (Italian, French) as well as two folk-linguistically very prominent dialects, the only Low-Alemannic variety in Switzerland of the city of Basel, and the dialect of the Wallis, a high-alemannic extreme case that is stereotypically considered unintelligible to people from outside the Wallis. Although the variety of bubbles associated to these categories is relatively wide, there is a certain preference for organic and/or floral shapes (high scores for bubbles 3, 4, 6 and 11).

It seems to me quite safe to assume that all entries located in the upper two quadrants of the plot are perceived as prototypical dialects/languages. In the case of the dialects we find categories covering the most important urban centers in the Swiss German area, whereas all dialects in the lower quadrants are somehow less prototypical and definitely not urban (one of the particularities of the canton of Aargau is precisely that it has no historical urban center, and stereotypically the saying goes that all the major towns of the canton simply belong to Zürich suburbia).

Table 4. Mappings and attributes for the St. Gallen dialect (cf. Table 2 above for the Bern dialect)

St. Gallen dialect	
bubbles	*attributes*
1 (10, 31%)	angular (2), bright, chiseled, clanking,
7 (8, 25%)	pointed, sharp (2), sharp edges, spiky
12 (3, 9%)	(2), spinose, straight, strident

For the remainder of this section, I propose to focus on two varieties of Swiss German, also located in clearly distinct sections of dimension 1 in Figure 3. The first of these two is again the Bern dialect, a variety spoken at the border of the Swiss German territories, close to the French part of the country. The second dialect is the St. Gallen dialect, spoken in the eastern part of German-speaking Switzerland, closer to the German and Austrian borders. St. Gallen dialect is close to Zürich and Standard High German in Figure 3. I will first give an account of the mappings of the St. Gallen di-

alect in order to be able to compare them to the mappings of the Bern dialect already discussed in the preceding section.

If one compares Table 2 to Table 4, it becomes clear that indeed there is a great amount of overlap between the St. Gallen stimuli and attributes and the stimuli and attributes chosen for Standard High German. As already noted above, it is thus important to differentiate between different dialects: not all of them trigger or carry the same mental "gestalts". The significant difference between Bern and St. Gallen dialects can be related to a traditional dislike of eastern Swiss varieties, a tendency which can be traced back several hundred years (Ris 1992: 756). Thus, it turns out that these internal Swiss German stereotypes have a very long life. And it is also important to note that not only people outside St. Gallen have these esthetically negative evaluations of the St. Gallen dialect; in our sample we also had speakers of this dialect who themselves exhibited exactly the same mapping pattern (cf. the groundbreaking study by Lambert et al. 1960 that revealed negative attitudes towards the participants' own varieties by using a matched guise design).

According to the sociolinguistic literature (Giles, Bourhis, and Davies 1975; Trudgill 1983), such esthetic and other judgments about varieties are mainly ideological constructions, based on either the cultural value attributed to the standard ("imposed norms hypothesis") or based on other social/cultural stereotypes ("social connotation hypothesis"). Despite the overwhelming evidence for the priority of sociological and cultural origin of evaluatory judgments and classifications of languages/varieties, dissident voices came out repeatedly advocating an intrinsic foundation especially of esthetic judgments ("inherent value hypothesis"). It is this third hypothesis that we will be examining in the remainder of this section.

The starting-point of the following observations was the perplexing finding that our participants were at the same time quite consistent regarding certain attributions and mappings and totally unable to give any other motivations for their attributes than linguistically/phonologically motivated ones. In the dialect imitation task (see section 2), many participants indeed imitated the St. Gallen dialect (or less specifically the so called "Ostschweizer Dialekte", eastern Swiss dialects) and overtly commented on the different vowel phonology compared to most other Swiss German dialects, as in the following transcript of the dialect imitation task:

Imitation by participant SH3 (native speaker of Bern dialect)
SH3: Well the easterners would say [*gives a mock St. Gallen example with*

*very open fronted /a/, /e/ and /i/ phonemes and a significantly raised voice
(f0)*]
INVESTIGATOR: do you always have to raise your voice here?
SH3:   [laughs]   no,   the   raising   of   the   voice   is,   now   –
I am imitating somebody who used this sentence […] you do automatically
– in order to say this, you go up, probably with the tongue, or, that is the…
INVESTIGATOR: you actually don't really open your mouth
SH3: yes you also have… I mean, this is very characteristic for this dialect,
you've got totally different – for many things you've got a totally different
position of the mouth

What the participant refers to here could be related to the actual phonologi-
cal makeup of typical St. Gallen dialect: the presence of, historically speak-
ing, a considerable amount of vowel raising and the absence of certain par-
ticular vowel lowerings. Figure 4 gives an account of these phonological
processes. The relevant processes are illustrated in Table 5.
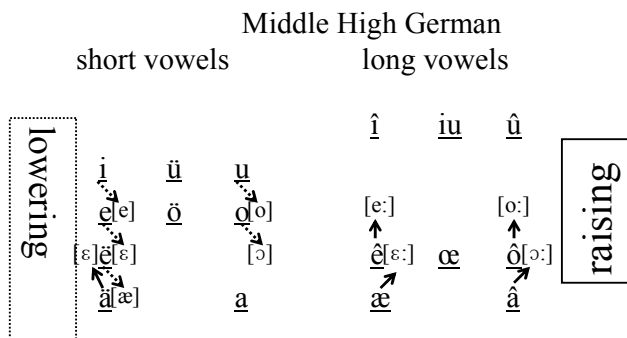
Figure 4. Raising and lowering in Swiss German dialects (cf. Haas 1978)

A salient consequence of the processes documented in Figure 4 and Table 5
is that any sample of authentic St. Gallen Swiss German speech contains
indeed significantly more high vowels compared to e.g. the Bern dialect.
As shown above, at least some particularly good "amateur linguists" among
our participants are actually aware of these differences – maybe not in
terms of phonological systems, but in terms of the articulatory differences
of the instantiations of the systems. Since these differences are salient and
systematic, we can at least try to ask the question whether there is any evi-
dence from research on sound symbolism (as it can be found in the early

gestalt psychology, Köhler 1933: 153) that allows to hypothesize that angular and pointed forms as in stimuli 1, 7 and 12 can be related to high vowels. Köhler's initial experiment presented his participants with different shapes drawn on a paper and the participants had to attribute artificial words such as *takete* and *maluma*. Unfortunately, these words not only vary the vowels, but also the consonants, which makes it impossible to tell whether there is a relationship between high vowels and chiseled, sharp and pointy forms.

Table 5. Raising and lowering in Bern and St. Gallen Swiss German and in Standard High German

Middle High German	Bern dialect	St. Gallen dialect	Standard High German	English translation or cognate
lowering:				
/i/:	[ʃlem]	[ʃlim]	[ʃlɪm]	'bad'
/u/:	[noɾ]	[nuɾ]	[nuːɐ̯]	'only'
/e/:	[blɛtəɾ]	[bletəɾ]	[blɛtɐ̯]	'leaves'
/o/	[ɔdəɾ]	[odəɾ]	[oːdɐ̯]	'or'
/ē/	[vælʃ]	[vɛlʃ]	[vɛlʃ]	'Welsh'
/æ/	[væːɾ]	[vɛːɾ]	[veːɐ̯]	'who'
raising:				
/ä/	[xæːs]	[xɛːs]	[kɛːzə]/[keːzə]	'cheese'
/â/	[ʃtrɑːs]	[ʃtrɔːs]	[ʃtʀaːsə]	'street'
/ê/	[sɛː]	[seː]	[zeː]	'sea'
/ô/	[sɔː]	[soː]	[zoː]	'so'

There is more recent research using a very similar experimental pattern. In their 2001 paper, Ramachandran and Hubbard report on a study which presented the participants with drawings such as the two given in Figure 5. The participants are requested to do the following task: "In Martian language, one of these two figures is a *booba* and the other is a *kiki*, try to guess which is which". The results of this study were clear-cut: between 95% and 98% of the participants assign *kiki* to the left hand shape and *booba* (or *bouba*) to the shape on the right hand side. Again, the problem is that the two words vary both consonants and vowels. Therefore I replicated the experiment with a total of 60 students at the universities of Lausanne and Fribourg (about one third Francophones, one third native speakers of Ger-

man and one third native speakers of other languages; none of these students had been involved in the project discussed above). I ran the experiment in two similar conditions, once opposing *wiiwii* vs. *waawaa* and once opposing *liilii* and *laalaa*. Both pairs hold the consonants constant and vary the vowels along the vertical dimension. The result again was perfectly non-ambiguous: between 90% and 100% of the students attribute the –ii– forms to the figure on the left and the –aa–forms to the figure on the right.[3]

Figure 5. The booba-kiki task by Ramachandran/Hubbard 2001

In view of the high consistency in the choice of the "spiky" visual stimuli chosen for the St. Gallen dialect, we can now hypothesize that there is more to the stereotype of the "sharp" and unpleasant nature of the St. Gallen dialect: there may well be a perceptual, synaesthetic correlate between the high frequency of occurrence of high vowels and the folk linguistic stereotype associated with the dialect. The differences between the phonological systems are perceptually salient, because they involve in many cases not the existence of totally unknown or inexistent phonemes in one or the other variety, but rather the contrastive distribution of the phonemes in the phonological system which thus leads to highly contrastive sound patterns in pairs of words, as exemplified in Table 5 above. It is these contrasts that are perceptually and thus folk linguistically salient, as Trudgill (1986: 19) notes:

> If differences between two accents involve simply the incidence of a particular phoneme in a given lexical set, then that difference will be very highly salient […]. English English Speakers are highly aware of US /æ/ in dance because they themselves have /æ/ in romance.

Despite the perceptual saliency of phonological contrasts between the two dialects discussed above that seems to be clearly supported, a third variety,

Standard High German, needs to be taken into account. As we have seen in Table 3 and Figure 3, Standard High German and St. Gallen dialect are often associated with similar visual stimuli. In terms of vowel phonology, the two varieties of German are quite similar, as shown in Table 5: In 7 out of the 10 variables listed, Standard High German has equally closed or even closer vowel qualities than St. Gallen German. In one of the remaining phonemes, MHG /i/, Standard High German is at least mid-way between the lower Bern variant and the higher St. Gallen variant. Given the fact that the visual stimuli match both the synaesthetic regularities discussed in this section as well as the cultural topoi discussed in section 2.2, we cannot entirely be sure whether the St. Gallen bubble choices are exclusively or at least mainly influenced by the vowel quality, or whether there is an indirect, secondary effect of the recognition of phonological proximity to Standard High German resulting in the choice of similar bubbles as for Standard High German. However, instances from the qualitative interview data such as the one discussed above repeatedly show that the folk spontaneously associate spikiness with the St. Gallen dialect, often explicitly mentioning the quality of the /i/ phoneme, which indeed is even more closed than the corresponding sound in Standard High German (cf. Table 5). Undoubtedly, a more refined methodology that would allow tapping even more directly into the folk's perceptional associations would be necessary to give a more confident answer to this question. For the time being, based on the evidence discussed above, it seems nevertheless not totally far-fetched to assume that there is a synaesthetic component that contributes to the quite consistent visual associations both in the case of the St. Gallen dialect and Standard High German. And, given the multifactorial nature of metalinguistic models and representations, there is no need to exclude more complex interactions such as a conspiracy of phonology, synaesthesia, and cultural models that, in a joint manner, reinforce this consistency.

Taking into consideration the data and analyses presented in this section, the inherent value hypothesis cannot be categorically rejected. This, however, obviously does not entail that sociolinguistic stereotypes are only perceptually grounded. But further research might show that there is a (universal?) esthetic or synaesthetic component to some of the most recurrent sociolinguistic stereotypes about languages or dialects.

3.   Conclusions

The goal of this contribution is to make three related points. Firstly, I have tried to show that a "cognitive turn" in the research on language attitudes and, more generally in the research on folk models of language, opens up new perspectives on old issues such as sociolinguistic prestige and esthetic evaluations of languages and varieties. Whereas standard sociolinguistics often keeps looking at the social value of particular varieties and variants in a monodimensional fashion (i.e. in terms of the high-low relationship), a cognitive study of the folk's mental models of languages and varieties reveals the multidimensionality and perspective-dependency of linguistic prestige. Hence, cognitive sociolinguistics is an adequate framework for the study of the cultural and mental models that set the stage for the sociolinguistic processes observed in empirical sociolinguistics. Secondly, I have tried to show that the use of visual stimulus material enables us to shed some new light on what I propose to call the "gestalt" representations of languages and varieties. The method presented in section 2 certainly can be improved: there is no doubt that it should and could be systematically developed and adapted to other sociolinguistic situations. The bubble task clearly bears a certain danger of circularity, since we do not know which ones of the attributes given by the participants are indeed related to a mental representation of the sound pattern of a particular variety and which ones might rather be triggered by the visual stimuli only. The actual reasons for the mappings from stimuli to varieties might be mysterious or more random than what the data discussed here suggest. Thus, the patterns discovered here must continue to be confronted with other evidence such as the one cited from Ris (1992) in section 2.3. Converging evidence also seems to emerge from Cuonz' (2008) study, where at least some of the same attributes as those described above are collected without any visual support whatsoever.

The evidence in section 2 seems to support the idea of consistent and recurrent visual correlates of mental models of languages and varieties. The third point was to give new support for the inherent value hypothesis that usually gets (prematurely) discarded by professional linguists – but not by the folk. So it may well be that what is believed to be a simple stanko, e.g. the "spiky" quality of certain varieties, turns out to have a solid and universal grounding in perception. But obviously, the evidence presented here cannot be more than a kick-off for a new and cognitively realistic consideration of these old questions.

Notes

1. Acknowledgement: Many thanks to Irmi Kaiser, Christina Cuonz and the reviewers of this volume for their very valuable comments on an earlier versions of this chapter.
2. Most of the data presented in this chapter have been collected in a Seminar on "folk linguistics" I have taught in 2004 at the University of Fribourg, Switzerland. Thanks to an extraordinary class of brilliant students who contributed to this research: Ine Baeyens, Michael Boller, Tom Baumann, Stefan Bosshart, Karin Brülhart, Cyril Cattin, Rahel Egli, Pamela Gerber, Sabrina Ghielmini, Sibyl Herrmann, Tobias Lambrecht, Guido U. Loosli, Rita Marty, Marina Petkova, Denise Pfammatter, Mirjeta Reci, Kristina Ruff, Kathrin Stadelmann, Franziska Suter, Jan Zenhäusern.
3. In my view, neither Köhler's original design nor any of the modified modern variants of this task can totally rule out possible mediating effects of the graphemes associated with the high and low vowels (e.g. <o> vs. <i>).

References

Berthele, Raphael
  2001   A tool, a bond, or a territory: language ideologies in the US and in Switzerland. In *LAUD Series A: General & Theoretical Papers*, 533.
  2006   Wie sieht das Berndeutsche so ungefähr aus? Über den Nutzen von Visualisierungen für die kognitive Laienlinguistik. In *Raumstrukturen im Alemannischen*, Hubert Klausmann (ed.), 163-176. Graz, Feldkirch: Neugebauer Verlag.
  2008   A nation is a territory with one culture and one language. The role of metaphorical folk models in language policy debates. In *Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems*, Gitte Kristiansen, and René Dirven (eds.), 301-331. Berlin/New York: Mouton de Gruyter.
Bourdieu, Pierre
  1982   *Ce que parler veut dire. L'économie des échanges linguistiques*. Paris: Librairie Arthème Fayard.
Christen, Helen
  1998   *Dialekt im Alltag. Eine empirische Untersuchung zur lokalen Komponente heutiger schweizerdeutscher Varietäten*. Tübingen: Niemeyer.

Collins, Allan, and Dedre Gentner
    1987   How people construct mental models. In *Cultural Models in Language and Thought*, Dorothy Holland, and Naomi Quinn (eds.), 243-265. Cambridge: Cambridge University Press.
Cuonz, Christina
    2008   *Stiefkinder der Linguistik: Ästhetische und affektive Sprachurteile in der linguistischen Biografie und Identität*. Paper presented at the Dritte Tagung Deutsche Sprachwissenschaft, Rome, Italy.
Ferguson, Charles A.
    1959   Diglossia. *Word* 15: 325-340.
    1991   Diglossia revisited. *Southwest Journal Linguistics* 10 (1): 214-234.
Geeraerts, Dirk
    2003   Cultural models of linguistic standardization. In *Cognitive Models in Language and Thought. Ideology, Metaphors and Meanings*, René Dirven, and Martin Pütz (eds.), 25-68. Berlin/New York: Mouton de Gruyter.
Giles, Howard
    1970   Evaluative reactions to accents. *Educational Review* 22: 211-227.
Giles, Howard, Richard T. Bourhis, and Ann Davies
    1975   Prestige speech styles: the imposed norm and inherent value hypotheses. In *Language in Anthropology. IV: Language in Many Ways*, William C. McCormack, and Stephen Wurm (eds.), 589-596. The Hague: Mouton.
Haas, Walter
    1978   *Sprachwandel und Sprachgeographie*. Wiesbaden: Frantz Steiner.
Häcki Buhofer, Annelies, and Harald Burger
    1998   *Wie Deutschschweizer Kinder Hochdeutsch lernen*. Stuttgart: Franz Steiner Verlag.
Hogg, Michael, Nicholas Joyce, and Dominic Abrams
    1984   Diglossia in Switzerland. A social identity analysis of speaker evaluations. *Journal of Language and Social Psychology* 3: 185−196.
Köhler, Wolfgang
    1929   *Gestalt Psychology*. New York: Liveright.
    1933   *Psychologische Probleme*. Berlin: J. Springer.
Kristiansen, Gitte
    2008   The group as a variable in the development of cognitive schemata. In *Body, Language and Mind, Vol. II*, CLR 35.2, Roslyn M. Frank, René Dirven, Tom Ziemke, and Enrique Bernárdez (eds.), 409-432. Berlin/New York: Mouton de Gruyter.

Kruskal, Joseph B., and Myron Wish
   1991   *Multidimensional Scaling*. Newbury Park/London.New Delhi: Sage
            Publications.
Labov, William
   1966   *The Social stratification of English in New York City*. Washington:
            Center for Applied Linguistics.
Lakoff, George, and Mark Johnson
   1980   *Metaphors We Live By*. Chicago/London: University Of Chicago
            Press.
Lakoff, George
   1987   *Women, Fire, and Dangerous Things: What Categories Reveal
            about the Mind*. Chicago/London: The University of Chicago Press.
Lambert, Wallace E., Richard C. Hodgson, Robert C. Gardner, and Samuel Fillen-
            baum
   1960   Evaluational Reactions to spoken languages. *Journal of Abnormal
            and Social Psychology* 60: 44-51.
Lasagabaster, David
   2005   Attitude/Einstellung. In *Sociolinguistics. Soziolinguistik. Ein inter-
            nationales Handbuch. 2. Edition 2. Volume*, Ulrich Ammon, Norbert
            Dittmar, Klaus J. Mattheier, and Peter Trudgill (eds.), 399-405. Ber-
            lin/New York: Mouton de Gruyter.
Le Page, Robert Breck, and Andrée Tabouret-Keller
   1985   *Acts of Identity. Creole-Based Approaches to Language and Ethnici-
            ty*. Cambridge, London.
Milroy, Lesley
   1992   *Language and Social Networks*. Oxford/Cambridge, MA: Blackwell.
Niedzielski, Nancy, and Dennis R. Preston
   1999   *Folk Linguistics*. Berlin/New York: Mouton de Gruyter.
Preston, Dennis
   2004   Language with an attitude. In *Handbook of Language Variation and
            Change*, John K. Chambers, Peter Trudgill, and Natalie Schilling-
            Estes (eds.), 40-66. Oxford/Malden: Blackwell.
Ramachandran, Vilayanur S., and Edward M. Hubbard
   2001   Synaesthesia - A Window into Perception, Thought and Language.
            Journal of Consciousness Studies 8 (12): 3-34.
Ris, Roland
   1992   Innerethnik der deutschen Schweiz. In *Handbuch der schweizeri-
            schen Volkskultur*, Paul Hugger (ed.), 749-766. Zürich: Offizin.
Siebenhaar, Beat
   2000   *Sprachvariation, Sprachwandel und Einstellung. Der Dialekt der
            Stadt Aarau in der Labilitätszone zwischen Zürcher und Berner
            Mundartraum*. Stuttgart: Franz Steiner Verlag.

Swiggers, Pierre
  1990   Ideology and the 'clarity' of French. In *Ideologies of Language*, John E. Joseph, and Talbot J. Taylor (eds.), 112-130. London: Routledge.
Trudgill, Peter
  1972   Sex, covert prestige, and linguistic change in the urban British English of Norwich. *Language in Society* 1: 179-196.
  1983   *On Dialect. Social and Geographical perspectives*. Oxford: Blackwell.
  1986   *Dialects in contact*. Oxford/New York: Blackwell.
Van Bezooijen, Renée
  1997   Normen met betrekking tot het Standaardnederlands. *Taal en Tongval, Themanummer 10 Standaardisering in Noord en Zuid*: 30-48.
Werlen, Iwar
  1985   Zur Einschätzung von schweizerdeutschen Dialekten. In *Probleme der schweizerischen Dialektologie*, Iwar Werlen (ed.), 195-266. Freiburg i.Ü.: Universitätsverlag.

A cognitive approach to quantitative sociolinguistic variation: evidence from th-fronting in Central Scotland[1]

Lynn Clark and Graeme Trousdale

Abstract

Language in use is central to both sociolinguistics and Cognitive Linguistics (CL). But despite the claim that CL is a usage-based model of language structure, attempts have only recently been made to synthesise aspects of CL and sociolinguistics (see, for example, the collection of papers in Kristiansen and Dirven 2008), and generally the treatment of this synthesis has been weak on empirical detail. Using techniques typical of modern (variationist) sociolinguistic data collection and analysis, we examine a phonological change in progress (TH-Fronting) in East-Central Scotland. We incorporate both 'social' and 'cognitive' factors in our analysis of this phonological change and suggest that such a unified approach not only provides a more parsimonious theory of language, but also enables us to account more comprehensively for the variability associated with this sound change. The outcome of such a synthesis is therefore beneficial to researchers in both sociolinguistics and Cognitive Linguistics.

Keywords: phonological change, th-fronting, usage, frequency, social, entrenchment, identity

1. Introduction

This chapter is concerned with the direct application of some key concepts in Cognitive Linguistics (cf. Langacker 1987, 1991; Taylor 2002; Croft and Cruse 2004) to patterns of variation in a community of speakers in central Scotland. Our main aim is to show how, by synthesising sociolinguistic methods of data collection and analysis with cognitive linguistic methods of interpretation, we can reach a deeper understanding of linguistic variation and change in a particular community. The chapter falls into three main parts.

In section 2, we begin by outlining some issues in linguistic cognition that are also reported to be operative in human social cognition. To a certain extent this is to be expected from general principles of Cognitive Linguistics (on which see Croft 2009); but the objective here is to draw on social psychological research to support our claim that categorization is operative in all types of cognition (specifically, both social and linguistic knowledge). This leads us to consider some other critical issues in Cognitive Linguistics, namely multiple inheritance, schematicity and full and partial sanction. Such concepts allow a fuller understanding of the nature of the sociolinguistic variation we have observed.

Section 3 consists of a brief description of the corpus and methods of collection, and a summary of the linguistic change in progress. This change, known as TH-Fronting, is well-documented in accent studies of British English (cf. Wells 1982, and the contributions to Foulkes and Docherty 1999b), but its spread into central Scots seems to be more complex than the changes affecting London English, for instance (on the latter, see further Tollfree 1999). This section provides a brief overview of the material collected and then a multivariate analysis of the data. This section also provides the data for the 'cognitive sociolinguistic' interpretation in the remainder of the chapter.

Section 4 is concerned with the implications for a usage-based model of linguistic and social cognition, situated particularly within the parameters of Cognitive Grammar, though clearly applicable to other models which form part of the cognitive linguistics enterprise. We argue that high frequency of use may explain some of the variability in TH-Fronting, but that frequency effects are only explicable when assumptions of a usage-based model are invoked**.**

Section 5 is the conclusion, where an approach to sociolinguistics that is sensitive both to aspects of formal linguistic theory and to the nature of social meaning is advocated.

2.   On the relationship between sociolinguistics and cognitive linguistics

An attempt to bring together (cognitive) linguistic theory with sociolinguistics necessitates a discussion of two related questions: first, how usage-based is cognitive linguistics; and second, how theoretical is sociolinguistics? The first question is relevant because while it is clear that sociolin-

guistics is firmly embedded in the analysis of language as it is used in its social context, it is less clear that this is true of cognitive linguistics, despite the various claims that it is a 'usage-based' model. We therefore concur with Geeraerts when he notes that "it is impossible to take seriously the claim that Cognitive Linguistics is a usage-based approach and at the same time to neglect the social aspects of language use" (Geeraerts 2001: 53). However, it is encouraging to note that recently there have been a number of publications within subdisciplines of cognitive linguistics that have addressed the issue of usage quite extensively, such as the following:

o   Word Grammar analysis of sociolinguistic variation in the morphosyntax of Buckie Scots (Hudson 2007b)
o   Construction Grammar analysis of the linguistics of (in)alienable possession in Lancashire English (Hollmann and Siewierska 2007)
o   Cognitive Grammar sketch of linguistic variation correlated with gender, sexuality and communities of practice (Watson 2006)
o   Cognitive Grammar discussion of lectal variation (Kristiansen 2003)
o   Discussion of usage-based methods in the context of Cognitive Linguistics (Tummers, Heylen, and Geeraerts 2005)

Sociolinguists, by contrast, have been accused of failing to pay close enough attention to developments in linguistic theory, and particularly, of failing to build a bridge between their research and that of general linguistics. Thus Cameron has noted "if sociolinguistics is to progress from description to explanation…it is obviously in need of a theory linking the 'linguistic' to the 'socio'." (Cameron 1997: 59). We believe that Cognitive Grammar is one such means of developing that theoretical link, but it is important to note that other accounts have also been attempted, including earlier work on Lexical Phonology and sociolinguistic variation in Tyneside English (Trousdale 2002), and Minimalist accounts of the morphosyntax of Buckie Scots (Adger and Smith 2005). However, both Trousdale (2002) and Adger and Smith (2005), while sensitive to both formal and sociolinguistic phenomena, provide an analysis which keeps the 'linguistic' very distinct from the 'social', in terms of modular knowledge. The research discussed here presents a more synthetic account.

   One area of linguistic research that has recently witnessed a weakening of the dividing line between formal linguistic theory and sociolinguistics is the exemplar based model of phonology. Exemplar based models of categorization suggest that large amounts of information are stored in episodic memory as 'exemplars' or collections of complex experiences, Such a

model has been readily applied to phonological variation by, for example, Docherty and Foulkes (2006) and Docherty (2008). While we are sympathetic to some versions of exemplar theory, we argue (as with Bybee 2001) that speakers not only store exemplars in episodic memory, but also that they abstract away from such exemplars to form a specific network in cognition, where a particular instance of a phoneme category serves as the prototype and others serve as context-specific extensions from that prototype. 'Context' refers not only to the linguistic context in which the sound appears, but also the (immediate and wider) social context in which the sound is used (Kristiansen 2006: 116). This is discussed in some detail in section 4.

One attractive aspect of exemplar theory for cognitive sociolinguistics is its application within the domain of social psychology. For instance, Rothbart and Lewis (1988) have shown that social categorization relies on humans' ability to perceive and rank more highly canonical over non-canonical exemplars in any given category. They argue that in cases of multiple inheritance, any individual will be stored or associated with a category to which they conform most clearly (e.g. a female rock climber would be categorized as 'athlete' more readily than as 'woman'), which extends to issues of attribute ranking. They also provide evidence to suggest that goodness of exemplar categorization is as prevalent in social categorization as it is in linguistic categorization. The critical issues which arise from such exemplar approaches to linguistic and social cognition are:

- that social and linguistic categories are formed using the same general principles (Hudson 1996)
- that multiple inheritance helps to explain how entities may be simultaneously instances of more than one category (Hudson 2007a and b)
- that peripheral members of both linguistic and social categories may still be accorded membership of the group on the basis of partial sanction (Langacker 1987, 1991).

In this regard, the account of phonological categories put forward by, for example, Mompean-Gonzalez, is equally applicable to sociolinguistic theory. He writes: "the phoneme category /t/, for example, could be conceived of as a category of sounds embedded in a wider network of knowledge structures from which the relevant attributes characterizing the category are drawn" (Mompean-Gonzalez 2004: 442), and such a wider network of structures may include sociolinguistic phenomena such as discourse context

and social stereotypes associated with particular modes of speaking and types of speakers (see Kristiansen 2006 for more on social and linguistic stereotypes). For instance, in the Scottish community investigated as part of the present research, it is clear from the following extract that, contained within this speaker's knowledge system of variants of the phoneme /Ө/, is some knowledge of the 'types' of speakers who use the labiodental variant [f]. For the speakers in this extract, the realisation of /Ө/ as [f] (in the word 'three' at least) is associated with a social type, namely 'mink' or 'tinky', terms used to describe undesirable and/or unintelligent people[2]:

(1a) **Scots**
    LC: only ever three?
    S: aye
    LC: would ye ever say free?
    S: noh cos then yer a mink
    LC: really?
    S: free, oh aye
    E: oh aye, tinky
    LC: ye hear folk saying it though eh?
    S: free aye it just makes ye sound stupit

(1b) **English translation**
    LC: only ever three?
    S: yes
    LC: would you ever say free?
    S: no because then you are a mink
    LC: really?
    S: free, oh yes
    E: oh yes, tinky
    LC: you hear people saying it though, don't you?
    S: free yes it just makes you sound stupid

In the remainder of the chapter, we make use of three particular concepts from Cognitive Grammar, namely 'entrenchment', 'schematicity' and 'full and partial sanction'. A brief definition of each of these is provided, in order to clarify how we are using these terms. A phonological form is said to be entrenched by frequency of successful use (on which see further Langacker 1987: 380); it is at this level at which exemplars are most clearly operative, as phonetic phenomena. While it is clear that forms become more entrenched as units through greater frequency, it is also clear that there is no arbitrary cut off point where a form is or is not entrenched –

such a process is a gradient phenomenon, where forms are located on a cline of entrenchment.

Speakers are able to abstract from such exemplars to form phonological categories which are more schematic. Thus speakers may observe a commonality across a set of phones, just as they may observe a commonality across a set of speakers. As the phoneme /t/ is schematic for the alveolar tap and the glottal stop (among other phones) in many varieties of British English, so the category 'British' is schematic for the category 'Weegie', 'Scouse' and 'Geordie'[3] (among other British social types). Newness in a system may be created by partial sanction, where an innovation shares only part of the specifications of its sanctioning schema. Thus, in the case of a dialect area in which only two instances of a given variable exist, say a dental and a glottal fricative, the introduction of a new labio-dental variant in a lexical set is an instance of partial sanction (where manner of articulation is a shared specification, place of articulation is not shared); in terms of social stereotyping, the concept of a male nurse is allowed for by partial sanction (human is the shared specification, male gender is not shared). Full sanction, at its most extreme, is what we presume to be the selection of the prototype of any given category.

3.   The data

3.1. Methods in data collection

The data presented here were collected from a group of 54 speakers who play together in two interrelated pipe bands in west Fife, Scotland, called West Fife High Pipe Band (hereafter WFHPB). The geographical location of WFHPB is shown in Figure 1. The data were collected by the first author over a period of 30 months using the ethnographic technique of long-term participant observation (Eckert 2000). Ethnography is the participation in the daily lives of a community over an extended period of time. It has recently been used as a research method in a number of sociolinguistic studies (e.g. see Mendoza-Denton 1997; Eckert 2000; Moore 2003) with the intention of understanding "the sociolinguistic dynamics of the community from the perspective of the community itself" (Wolfram and Schilling-Estes 1996: 106). Crucially, this is often coupled with an analysis of the

social practices of a particular group of speakers, rather than a focus on the hierarchical social structure of the group, with the explicit intention of understanding the ways in which speakers imbue linguistic variation with social meaning.

Figure 1. Map showing the location of west Fife

The sociolinguistic interview has been the most common method of data collection among variationist sociolinguists (Milroy and Gordon 2003: 61). This is typically a one-on-one exchange between the researcher and the informant although variations of this include interviewing two or more speakers together (e.g. Labov 1972). The goal of the sociolinguistic interview is to elicit 'natural' or 'casual' speech and it has been used as a technique to reduce the effects of the 'observers paradox' i.e. the problem of observing how people speak when they are not being observed (Milroy and Gordon 2003: 49). Milroy and Gordon explain that the interview is a "clearly defined and quite common speech event to which a formal speech style is appropriate" (2003: 61) and that it generally involves interaction

between strangers. It is therefore inappropriate to couple the sociolinguistic interview with participant observation.

The informants were not questioned using a structured interview. Instead, the conversations that comprise the majority of the corpus were collected in the summer of 2006, roughly 2 years into the participant observation, and centered on a sorting task that the informants were asked to complete in small groups of friends[4]. Consequently, no two recordings are the same in this study. This does not necessarily mean, however, that the data from these interviews are not comparable. It is "unlikely that any two interviews will be the same no matter how structured the approach of the researcher" (Moore 2003: 45). The resulting data consists of 38 hours of recorded speech which have been fully transcribed and amounts to a corpus of 360,000 words.

3.2. TH-Fronting in Scotland

Although Wells invokes 'TH-Fronting' to refer to 'the replacement of the dental fricatives [Θ, ð] with the labiodentals [f] and [v] respectively' (Wells 1982: 328), we follow Stuart-Smith and Timmins (2006) who adopt the term only with reference to the voiceless variants (because the voiced and voiceless variants of (th) pattern differently in Scottish English). The first reported evidence of TH-Fronting in Scotland is given by Macafee (1983: 54) as occasional and sporadic but the main body of research on TH-Fronting in Scotland comes from the analysis of two corpora collected in 1997 and 2003, both of which form part of a large research project into language variation and change in Glasgow (Stuart-Smith and Tweedie 2000). The spread of TH-Fronting has also recently been investigated in the New Town of Livingston (Robinson 2005) which is situated approximately 15 miles from Edinburgh and 30 miles from Glasgow.

Accounts of TH-Fronting across Britain (e.g. Williams and Kerswill 1999; Kerswill 2003; Stuart-Smith and Timmins 2006) have correlated this linguistic change in progress with 'macro' social factors such as age, sex and social class and have typically found that this is a change that is being led by working class, adolescent males. The data for (th) in WFHPB, when stratified by age and sex[5], are charted in Figure 2.

Although there are no speakers in the corpus older than 42 years old, real time data from the Linguistic Atlas of Scotland (Mather and Speitel 1986) suggests that the labiodental variant is not a traditional feature of this

dialect area and so it is possible to rule out the effects of age-grading in the data. The general trend in these data follows the pattern of other studies of (th) in Britain – this is a change in progress that is being led by younger adolescents. Older adolescents and adults favor the standard dental variant and younger adolescents favor the incoming labiodental variant. In comparison with other research on this variable, there is, however, one significant difference: unlike in most other communities, it seems that it is females who are the highest users of the labiodental variant in WFHPB[6]. The mean difference between use of (th): [f] among males and females in this corpus is significant (t-test: t=5.766, df. = 53, p< 0.0001). There are no females in the 25+ year old age bracket in WFHPB but in the 12-15 year olds and the 16-24 year olds, females are using higher proportions of the labiodental variant than males in this community.
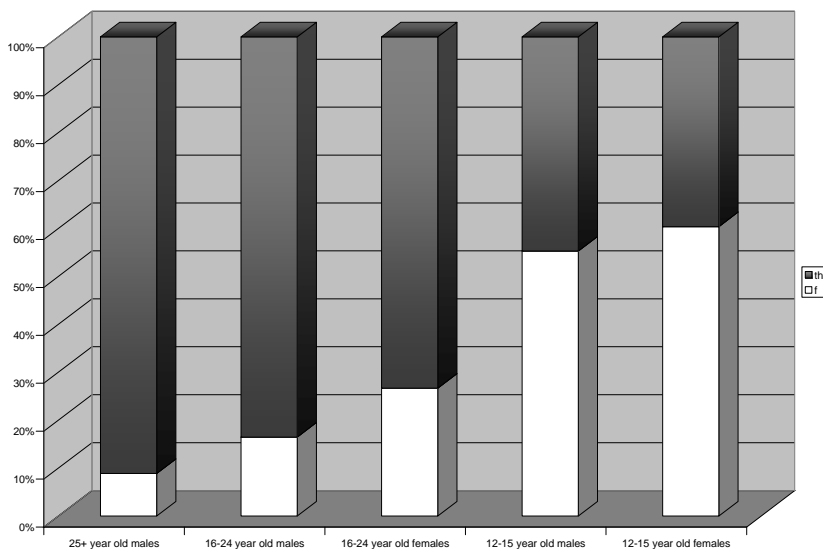

Figure 2. Variants of (th) in WFHPB stratified by age and sex

In order to reach a better understanding of the patterning of (th) in this community, we felt it necessary to consider the potential effects of a number of other (social, linguistic and cognitive) constraints on variation. To do this, it was necessary to employ more advanced statistical techniques.

3.3. Variable Rule Analysis

Tummers, Heylen, and Geeraerts (2005: 246) suggest that "in spite of its adherence to a maximalistic and usage-based theoretical framework, the use of advanced statistical techniques as yet turns out to be relatively sparse [in CL] and subordinate to the theoretical objectives". And yet, sociolinguistics has a well established tradition of employing advanced statistical techniques in the empirical analysis of usage data.

Variable rule analysis, or varbrul, is a statistical device that is commonly used in sociolinguistics to ascertain the effects of various independent factors influencing the distribution of a dependent variable by means of stepwise multiple regression[7]. It does this by computing the effect of one independent variable (or factor group) while explicitly controlling for the effects of all other known independent variables (Guy 1993: 237). Varbrul requires discrete variants for both the dependent and independent variables (or factor groups) and so the researcher must code each factor group (which contains a number of factors) in this way. The 'linguistic' factor groups coded in this varbrul analysis are provided in Table 1 and the 'social' factor groups are in Table 2.

Table 1. Linguistic factor groups for varbrul analysis of (th)

Linguistic Factor Group	Factors
Preceding phon. segment	Front vowel / Back vowel / Front consonant / Back consonant / Pause
Following phon. segment	Front vowel / Back vowel / Front consonant / Back consonant / Pause
Preceding #	Present / Absent
Following #	Present / Absent
Preceding [f]	Present / Absent
Place of (th) (syllable)	Onset / Coda
Place of (th) (word)	Initial / Medial / Final
Lexical category	Place names & proper names / Ordinals / Other
Lexical frequency	Low / Low-mid / High-mid / High

Table 2. Social factor groups for varbrul analysis of (th)

Factor Group	Factors
Individual speaker	54 individual factors, one for each speaker
Speaker sex	Male
	Female
Friendship group membership	A "They act hard all the time"/ "fancy tune folk"
	B "Tiny wee pipers"
	C "The new folk"
	D "Pipe band geeks"/ "Ex-Dream Valley"
	E "comedians" /"Same dress sense, same music taste, same easy going attitude"
	F "Fun/up for a laugh, not very serious"
	G "that's a fake ID son"
	H "senior drummers"/"pipe band geeks"
	I "one big happy family"
	J "On the fringe"
	K "13 goin on 30"
	L "goths"/ "new lassie pipers"
	M "Lazy PPl!"
	N "Dollar lassies"
	O "Under agers"
	P "Novice tenor section 'WILD'!!"
	Q No CofP affiliation
Age	12-15 years old
	16-24 years old
	25+ years old
Length of time in the band	< 10% of age
	10-19% of age
	20-29% of age
	30-39% of age
	40-49% of age
	50+% of age
Area of residence	Lochgelly
	Balingary
	Lochore
	Cardenden
	Cowdenbeath
	Falkland
	Glenrothes
	Scotlandwell
	Rosyth
	Dunfermline
	Burntisland
	Dollar
	Leven
	Dundee
	Crossgates

Many of the social and linguistic factor groups were included to test pervious findings from the research literature on th-fronting in Britain (e.g. speaker age, speaker sex, phonological context of variation). These factor groups are therefore typical of the type often found in sociophonetics and require no further comment here. There were, however, two main differences between our approach and most other mainstream analyses of phonological variation and change; these were (a) the methods used in order to reach the social factors presented within the factor group 'friendship group/community of practice membership' and (b) the inclusion of lexical frequency as a 'cognitive factor group' in the analysis of variation.

The friendship groups presented in Table 2 are based on a combination of observations made during the 2-year period of ethnographic fieldwork and a sorting task that the members of WFHPB were asked to carry out. This was modeled on a sorting task developed by Mathews (2005) in her research on the category labels that were given to adolescent girls in an American high school. Each informant was presented with a group of cards; each card contained the name and/or nickname of every other member of WFHPB. The informants were then asked to group the cards together into friendship groups, place these groups of cards inside an envelope and label the envelope with something that they felt characterized the behavior of the group. The results of each separate sorting task can also be collated into an aggregate matrix. It is then possible to find sub-groups (or cliques) within this aggregate matrix by employing the clique analysis built into social network software packages such as UCI NET (Borgatti, Everett and Freeman 2002). The results of the clique analysis suggest that the groups presented as friends and/or communities of practice in Table 2 are the most salient or most clearly identifiable social groups in WFHPB[8].

In order to discover whether there is a significant correlation between lexical frequency and th-fronting in WFHPB, it is first necessary to consider how best to measure lexical frequency in these data. This is problematic because, as Bybee explains, "there is no one method for doing frequency research" (Bybee 2007: 16). Often researchers interested in frequency effects take the frequency value of a particular lexical item from a list of frequency counts such as that provided by Baayen, Piepenbrock, and Gulikers (1995) in the form of the CELEX lexical database (employed by Hay 2001). However, the highly vernacular nature of the WFHPB data meant that a number of potential th-fronting sites occurred in local placenames, nicknames and other non-standard lexical items, all of which appear much more frequently in the WFHPB corpus than in any corpus of British Eng-

lish   (BNC   http://www.natcorp.ox.ac.uk/   or
http://www.scottishcorpus.ac.uk/ for Scottish English).

We therefore advocate an approach which measures lexical frequency
relative to the local corpus from which the data were collected[9].

Because varbrul requires discrete variants of all variables, it was unfor-
tunately necessary to convert the continuous measurement of lexical fre-
quency into discrete categories. Rather than create arbitrary cut points in
the data or force category divisions in order that the number of tokens in
each was approximately equal, the raw results for (th): [f] in all variable
lexical items[10] were plotted against lexical frequency in a scattergram and
natural 'bunches' in the data were highlighted (see Figure 3). While these
categories do not contain an equal number of tokens or types, they
represent the frequency categories that naturally emerged from the data and
so these were used to quantify the continuous measurement of lexical fre-
quency into a categorical format for varbrul.

In order to achieve a valid varbrul analysis, the factor groups must be
'orthogonal' (Guy 1988: 136) i.e. there must be minimal overlap between
the factor groups. This can often be difficult to achieve, for example in the
'linguistic' factors coded here, there is a certain amount of overlap between
the factor groups 'place of (th) in the syllable', 'place of (th) in the word'
and 'word boundary'. Interactions (or associations) between social factor
groups is perhaps even more difficult to avoid as there is more potential for
overlap (see Bayley 2002: 131): individuals tend to form friendship cliques
with others of the same sex, of roughly the same age and from the same
local area. It is extremely important to consider the effect of these distribu-
tional interactions when conducting statistical analyses. In Varbrul, it is
possible to spot such interactions with the 'crosstabs' function because the
cells of a crosstabulation will be unevenly occupied when there are interac-
tions between factor groups. We attempted to tease apart any possible inte-
ractions between different factors influencing variation by running the
analysis repeatedly and including different factor groups in the analysis
each time. We then compared the results of each analysis using a likelihood
ratio test to find which provided the best 'fit' and therefore the best indica-
tion of the likely factors influencing this variation[11]. Table 3 is organized to
show the factor groups in the order of their significance on the variation.
Factor groups not selected as significant are not shown in this table.

Table 3. Multivariate analysis of the contribution of factors selected as significant
to the probability of (th): [f]

	Factor weight	% of (th): [f]	N
Friendship group membership			
A "They act hard all the time"/ "fancy tune folk"	0.71	67	49
B "Tiny wee pipers"	0.95	93	56
C "The new folk"	0.89	85	59
D "Pipe band geeks"/ "Ex-Dream Valley"	0.10	7	27
E "comedians" /"Same dress sense etc."	0.32	32	28
F "Fun/up for a laugh, not very serious"	0.75	75	24
G "that's a fake ID son"	0.58	59	34
H "senior drummers"/"pipe band geeks"	0.09	9	76
I "one big happy family"	0.45	45	20
J "On the fringe"	0.21	23	57
K "13 goin on 30"	0.60	59	39
L "goths"/ "new lassie pipers"	0.51	55	87
M "Lazy PPl!"	0.31	30	78
O "Under agers"	0.48	44	32
P "Novice tenor section 'WILD'!!"	0.79	78	45
Q No CofP affiliation	0.35	34	73
Range	86		
Preceding [f] in the word			
Preceding [f]	0.81	68	22
No preceding [f]	0.49	48	762
Range	32		
Syllable structure/place of (th) in the word			
(th) in onset position/word initially	0.58	55	298
(th) in coda position/word finally	0.37	38	486
Range	21		
Type of lexical item			
Place names and proper names	0.42	48	351
Ordinals	0.42	39	324
All other lexical items	0.61	53	109
Range	19		
Frequency of lexical item			
Low frequency	0.41	39	242
Low-Mid frequency	0.47	57	148
High-Mid frequency	0.53	60	139
High frequency	0.58	48	255
Range	17		

Corrected mean 0.52, Log Likelihood -401.980, Total N 784

The corrected mean (also known as the input value) is a measure of the rate of 'rule application' or "an average frequency of occurrence of the application value of the dependent variable" (Paolillo 2002: 76). A corrected mean of 0.48 means that, all other things being equal, the likelihood of (th) being realised as [f] in this corpus is around 48%. The log likelihood value measures the likelihood that a particular set of data has been generated by the model. This is the value used when considering which combination of factors provides the best 'fit' of the model to the data. The total N is simply the total number of tokens included in the final run of the analysis. Finally, the factor weight is a value that is assigned to each factor during the analysis. It is essentially a measure of relative influence on variation. The number ranges from 0 to 1 and it is often stated in the literature that a factor weight of greater than 0.5 favours the application value (in this case, the labiodental variant) and a weight of less than 0.5 disfavours the application value).

The results of the varbrul analysis show that lexical frequency is the last significant factor group to remain in the analysis. In other words, of all the factors influencing variation in these data, lexical frequency has the weakest effect, while still remaining significant. We know that this is the case because, of the factor groups to achieve significance, this factor group shows the smallest range between the highest and lowest factor weights. The next most important constraint on the variation is syllable structure/place of (th) in the word. The results for this factor group suggest that when (th) occurs in syllable/word[12] initial position, it favors the dental fricative and when it occurs syllable/word finally, the labiodental is more likely to occur. The next most important constraint on the variation in (th) is lexical category. These results support the proposition made by Stuart-Smith and Timmins (2006) that ordinals and place names may be more resistant to the spread of TH-Fronting than other lexical items as ordinals, place names and proper names favor retention of the dental fricative. The next most significant constraint on variation in (th) is the factor group which codes for a priming effect. If the lexical item in question contains an [f] somewhere before the variable (such as in *fourth*) then the variable strongly favors the labiodental variant. Finally, the factor group 'friendship group membership' substantially outranks all other constraints on the variation. There is a very strong correlation between the use of the labiodental fricative and membership in a particular social group in this community.

The main advantage to employing a varbrul analysis is that it provides the capability to model both social and linguistic factors simultaneously

impacting on a speaker's (or group of speakers') choice of variants and to rank their relative strength and significance. However, as Guy warns, statistical analysis does not in itself explain variability: "Varbrul only performs mathematical manipulations on a data set. It does not tell us what the numbers mean, let alone do linguistics for us" (1988: 133). In order to interpret these findings, it is necessary to contextualize them within a wider theoretical framework. The remainder of the chapter will consider to what extent it is possible and beneficial to interpret some of these findings within a cognitive framework.

4. Cognitive Analysis of (th).

4.1. Lexical frequency as a 'cognitive' factor group

In some respects, the first step towards a cognitive analysis of (th) has already been taken by including factor groups such as 'lexical frequency' in the varbrul analysis because, in order to be able to explain these results, it is necessary to adopt certain theoretical assumptions of a usage-based model of language structure[13].

Perhaps the defining feature of a usage-based model is that there is assumed to be an unquestionable relationship between language structure and language use; language use is involved in shaping the grammar of individual speakers. It is argued that speakers' linguistic systems are grounded in 'usage events' or instances of producing and understanding language and that these are the bases on which the linguistic system is formed.

Kemmer and Barlow (2000: ix) describe the relationship between language structure and language use as a 'feedback loop' since experience of language both results from and also continues to shape the speaker's linguistic system. This necessarily implies that the frequency with which different parts of the language system are used will also affect the way in which the system is organized and stored in cognition: "since frequency of a particular usage-pattern is both a result and a shaping force in the system, frequency has an indispensable role in any explanatory account of language" (Kemmer and Barlow 2000: x).

In Cognitive Grammar, the relationship between frequency of use and the 'entrenchment' of the linguistic unit is particularly important (Langack-

er 1987: 59); for Langacker, entrenchment is the result of frequency of successful use[14]. As Langacker explains, the occurrence of any type of cognitive activity leaves behind a trace in cognition and the more that this type of activity recurs, the more entrenched the trace will become in cognition. Cognitive Grammar assumes (based on evidence from cognitive psychology) that all aspects of cognition are organized in the same way – as a cognitive network. In this model of the mind, nodes are linked in a pair-wise relationship by arcs. Nodes can correspond to any kind of linguistic structure (semantic, phonological or symbolic) and the arcs which link these nodes are categorization relationships. In any given usage event, a speakers and hearers will activate the particular nodes (at both the phonological and semantic pole) that correspond with the meaning that they are trying to convey or comprehend. As a particular node in the cognitive network is activated, it becomes more entrenched, which leads to the probability that it will be reselected.

In research on frequency effects in language change, the direct relationship between entrenchment and lexical frequency has led to the generalization that high frequency words and phrases are more entrenched in cognition (or have stronger lexical strength) and are therefore more easily accessed and are less likely to undergo analogical change. By contrast, low-frequency items are less well entrenched and so they are often difficult to access and more susceptible to analogical change (Bybee 2001: 28-9).

It is therefore clear that by adopting certain key theoretical assumptions of the usage-based approach, the frequency effect patterns we see in language begin to be explicable. Without adopting, at the very least, some recognition of the existence of a correlation between lexical frequency and cognitive entrenchment, it is very difficult to move beyond simply a descriptive account of frequency effects in language.

4.2. The relationship between linguistic variation and social meaning

Another potential area of crossover between the disciplines of sociolinguistics and Cognitive Linguistics lies in understanding the relationship between social and linguistic categories in the cognitive network (see also Kristiansen 2006: 108). This is exactly the task that is occupying many sociolinguists at the moment – they are becoming increasingly interested in the relationship between linguistic variation and social meaning at a very local level and asking the question "how do variables mean?" (Eckert 2002:

4). This is important because the sound change we explore in this article (TH-Fronting) is associated with particular social meanings for the highly-localized group from whom the data were collected.

TH-Fronting is one of a number of consonantal changes taking place in non-standard varieties of British English that are showing rapid changes, allegedly in the direction of a Southern English model of pronunciation. In Scotland, this has led to a number of media speculations that Scottish youngsters are being "influenced by Frank Butcher and other Cockneys in *Eastenders*" (Daily Record, 27th June 2000; cited in Stuart-Smith, Timmins, and Tweedie 2007: 221).   However in this community (as in many working class communities across Scotland) there remains a strong anti-English attitude:

(2a) **Scots**
     S: English. Well see the English maist folk'll just hink eh pricks
     LC: aye?
     S: aye

(2b) **English**
     S: English. Well see the English maist people will just think of pricks
     LC: yes?
     S: yes

It therefore seems unlikely that the speakers who are TH-Fronting in this community are consciously accommodating towards a linguistic variety that they associate with an 'English' stereotype. In fact some speakers seem to regard it as a local feature[15]:

(3a) **Scots**
     LC: dae you associate it wi English folk?
     N: what 'thanks' wi a-an 'f'?
     LC: mm
     N: nuh a wid eh said it wiz em, nah more a Fife way eh saying it
     LC: dae ye hink so?
     N: or a lazy way
     LC: a lazy way
     N: //but a-a've never thought English folk said it

(3b) **English**
     LC: do you associate it with English people?
     N: what 'thanks' with a-an 'f'?

LC: mm
N: no I would have said it was em, no more of a Fife way of saying it
LC: do you think so?
N: or a lazy way
LC: a lazy way
N: //but I-I've never thought English people said it

Also, because TH-Fronting (as with other consonantal changes including l-vocalisation, t-glottaling and the use of labiodental /r/) is being driven by adolescents, this has led several researchers (e.g. Foulkes and Docherty 1999a: 15, Milroy and Gordon 2003: 134) to suggest that these changes may represent a new set of 'youth norms' that are associated with 'trendy and hip London lifestyles' and 'youth culture' (Dyer 2002: 108). These 'youth norms', although originating in the south of England, are no longer associated with geographical or regional space but arguably exist in cultural or ideological space (i.e. 'spatiality', see Britain 2002) and represent a set of features which adolescents can orientate towards (Anderson 2000, cited in Stuart-Smith, Timmins, and Tweedie 2007). However, not all younger speakers are behaving in the same way in this regard. In other words, simply labelling this feature as a 'youth norm' tells us nothing about the *type* of youths who are driving this change forward.

When the speakers themselves were asked how they view the variation in TH-Fronting, many of them found it difficult to articulate their metalinguistic awareness, however the term 'rough' was used repeatedly to describe both TH-Fronting and other 'bad Scots'[16] dialect features:

(4a) **Scots**
    N: the wans that have been in the band for a while or a couple e
    years anyway an they talk rough like
    LC: what dae ye mean talk rough?
    N: [laughing] a know it's guid coming fae me but they talk real rough
    [laughs]

(4b) **English**
    N: the ones that have been in the band for a while or a couple o
    years anyway, and they talk rough like
    LC: what do you mean 'talk rough'?
    N: [laughing] I know it's good coming from me but they talk real rough
    [laughs]

It seems reasonable to suggest then that in this community, TH-Fronting is perceived as 'meaning' a number of different things. Some speakers view it as a local (Fife) feature, some associate it more with younger speakers and to some it indexes 'roughness'. It is no surprise then to find that the groups of speakers which favour the labiodental variant in the multivariate analysis are those which are perceived by others in the band as youthful and/or 'rough' (see Table 4).

Table 4. Friendship groups that strongly favor the labiodental fricative

	Factor weight	% of (th): [f]	N
B "Tiny wee pipers" (youth/inexperienced)	0.95	93	56
C "The new folk" (youth/inexperienced)	0.89	85	59
F "Fun/up for a laugh, not very serious" (youth)	0.75	75	24
P "Novice tenor section 'WILD'!!" (rough)	0.79	78	45
A "They act hard all the time"/ "fancy tune folk" (rough/experienced)	0.71	67	49

Table 5. Friendship groups that strongly disfavor the labiodental fricative

	Factor weight	% of (th): [f]	N
D "Pipe band geeks"/ "Ex-Dream Valley"	0.10	7	27
H "Senior drummers"/ "Pipe band geeks"	0.09	9	76

The 'tiny wee pipers' and 'the new folk' are two friendship groups that consist of the youngest and most inexperienced pipers in the band. The 'tiny wee pipers' are between 12-13 years old and none of them have yet played with the band in competition[17]. Because of their inexperience, they are often asked to sit to the side when the band is practicing and so they have formed their own friendship clique by default. 'The new folk' are a little older (13-14) but have been in the band less than a year and are still on the periphery. The group labelled 'fun, up for a laugh, not very serious' consists of three men who range in age from 20-33 but who typically enjoy living a bachelor lifestyle. The group labelled 'Novice tenor section WILD!!' is a group of four 13 year old girls who play the tenor drum in the band and who are best friends both in and out of band (two of them are twin sisters). The majority of their conversations consist of stories of un-

derage drinking, boyfriends and trouble with teachers and parents. Finally, the group labelled "They act hard all the time"/ "fancy tune folk" is a group of 4 teenage boys aged between 12 and15. They tend to wear expensive tracksuits and football tops and they project a confident, self assured, macho image.

Unsurprisingly, the groups which strongly disfavor the labiodental variant are characterized by opposite qualities to those that favor the labiodental variant: maturity, seniority and 'geekieness' (see Table 5). The individuals that comprise the group labelled "Pipe band geeks"/ "Ex-Dream Valley" are all in their mid to late twenties and have all played together either in this pipe band or in a related local band for a number of years. At a typical pipe band practice at WFHPB, these three individuals are immediately discernible. They all have 'professional' jobs (one is a banker, one works in local government and one is a police officer) and the clothing that they choose to wear at the pipe band practice reflects this aspect of their identity. This is expressed more explicitly by Lucy in the following extract as she describes one of the members of this friendship group:

(5a) **Scots**
L: he comes in fae work a lot an he's got his shirt an tie on, a hink people sort eh look doon at him because...he's got like a sortae upper class job sortae hing wi the suit sortae hing

(5b) **English**
L: he comes in from work a lot an he's got his shirt and tie on, I hink people sort of look down at him because...he's got like a sort of upper class job sort of thing with the suit, sort of thing

The group labelled "Senior drummers"/ "Pipe band geeks" have known each other for a number of years and they also all hold professional positions both inside and outside the pipe band (they are all involved in teaching within the band). WFHPB travel to and from competitions on a hired bus. The back seats of the bus are typically noisy and unruly with lots of alcohol, smoking and singing. In the following extract, Pete explains that the no-nonsense attitude of this group of friends is typified by the place that they occupy on the band bus:

(6a) **Scots**
P: at the front ye normally get me, Lewis, Connor...
LC: right

P: guys that just like tae travel there, read their paper, have a laugh aboot what wiz on the telly

LC: aye

P: an talk aboot the fitbaw an stuff like that like Bobby used tae be on oor bus fae Dream Valley, we used tae have a quiz an stuff like that in the mornings an stuff like that eh. That's what it wiz that's-that's the wiy we did it. A dinnae mind

(6b) **English**

P: at the front you normally get me, Lewis, Connor...

LC: right

P: guys that just like to travel there, read their paper, have a laugh about what was on the television

LC: yes

P: and talk about the football and stuff like that, like Bobby used to be on our bus from Dream Valley, we used to have a quiz and stuff like that in the mornings and stuff like that. That's what it was that's-that's the way we did it. I don't mind.

The argument presented here therefore follows Hudson (1996: 246) and suggests that speakers, as agents, can *choose* to signal their affiliation with a particular social group or type by selecting the linguistic variants that they associate with that social type. Speakers typically feel more or less affinity with a given social type but the linguistic choices available to signal allegiances are often binary (e.g. (th): [Ө]~[f]). Speakers can therefore use frequency of linguistic variation as a signal of their degree of allegiance to the social types that they perceive to be linked to each variant (Hudson 1996: 248). In other words, speakers can use linguistic variation as an *'act of identity'* (Le Page and Tabouret-Keller 1985). In Cognitive Grammar terms, speakers abstract over salient displays of style (such as dress, behavior *and* speech) and create schematic categories. Crucially, this process of abstraction occurs over both social and linguistic knowledge and so speakers create schematic categories at the phonological pole (e.g. the phoneme /Ө/) and the semantic/meaning pole (e.g. 'tiny wee pipers'). Because Cognitive Linguistic models (and Cognitive Grammar in particular) assume a non-modular view of language, this allows connections to be made in the mind between these categories of social and linguistic knowledge. The repeated co-activation and entrenchment of particular (social and linguistic) nodes and links in the cognitive network enables each speaker to associate social knowledge with particular linguistic variants. Speakers can then choose to signal their affiliation with a particular social type by activating

the linguistic variants that they associate with a given type. For instance, speakers in group D, F and H (see Tables 4 and 5 for details) may have, in their mind, an association with the realization of (th): [f] and younger speakers. Speakers in group F can therefore project their youthful image by using a higher frequency of the labiodental variant while speakers in group D and H can avoid using this variant for the same reasons i.e. to project a more 'grown-up', responsible image.

In describing the cognitive processes involved in linguistic variation, we have consistently emphasized the 'choice' that is available to the speaker. We acknowledge, however, that the individual is not without constraint - language is not "an identity 'free-for-all', a dressing-up box from which we can freely pick whatever suits us at that moment" (Britain and Matsumoto, 2005: 14). The internal structures of the language also impose considerable constraints on the scope for marking social type allegiances, such as those considered in the varbrul analysis above. Perhaps the truth of the matter then lies somewhere between the two extremes - while speakers can use linguistic variation to signal their identity, they may be "unconsciously motivated to claim symbolic identity with a reference group" (Labov 2002: 9).

The assumption so far has been that linguistic choices index social categories directly and while this may be the case in the mind of the individual speaker, it is important to note that the precise social meaning of TH-Fronting in a community is not the same for every speaker. This point has also been noticed recently by sociolinguists:

> The bottom line is that more than one group of speakers can use the same variable – but differences in the practices of these speakers will imbue that *same* variable with *different* meanings. (Moore 2003: 11, *original emphasis*).

Again, this sociolinguistic fact is entirely predictable within a usage-based framework: the linguistic structure that is abstracted is largely determined by a speaker's previous experience (Langacker 1987: 380), and as no two speakers will have had exactly the same linguistic experiences, each speaker will abstract a (minimally) different linguistic structure. The precise social meaning of TH-Fronting for each speaker is therefore dependent on a variety of factors including the type of speakers that each individual has experienced using the variant and the context in which it was used.

5. Conclusion

This article has argued, in line with the view adopted by other contributions to this volume, that Cognitive Linguistics and sociolinguistics are not mutually exclusive disciplines and, moreover, that by adopting some of the theoretical assumptions of a usage-based model, we can reach a more explanatory account of the relationship between linguistic and social variation in the mind of the individual speaker. We have demonstrated that this is possible in three main ways:

1. By including 'cognitive' factor groups into a traditional multivariate analysis and interpreting the statistical results through invoking assumptions of usage-based models of language structure.

2. By invoking a non-modularist, network approach to meaning (including social meaning) and beginning to understand the relationship that exists in the mind of the speaker between social meaning and linguistic variation.

3. By accepting the principle of multiple inheritance and beginning to understand how the same linguistic variable can have a range of different social meanings within the same community.

Notes

1. We are grateful to Dick Hudson and Kevin Watson for their invaluable comments on an earlier draft of this paper. We would also like to acknowledge the helpful comments from the organizers and the audience at the session on Cognitive Sociolinguistics at ICLC (2007) where this work was originally presented. Finally, we of course thank the two anonymous reviewers and the editors of this volume for their comments and suggestions for improvement.

2. Where excerpts are given in Scots, the English translation is provided immediately following. This applied to all extracts in the article.

3. 'Weegie', 'Scouse' and 'Geordie' are terms used to describe natives of the cities of Glasgow, Liverpool and Newcastle upon Tyne respectively.

4. See section 3.2 for a fuller discussion of this as a method of data collection.

5. WFHPB is socially fairly homogeneous and so it is not possible to stratify the speakers by social class.

6. Females also showed higher proportions of TH-Fronting in the Glasgow corpus collected in 1997 but in the corpus collected in 2003, males were using the labiodental variant more than females (see Stuart-Smith and Timmins 2006).

7. GOLDVARB X (Sankoff, Tagliamonte, and Smith 2005) was used to perform the varbrul analysis in this research.

8. One reviewer questioned the extent to which the friendship groups presented in Table 2 could be said to cohere linguistically. The internal coherence of friendship groups/communities of practice is an issue that deserves far greater attention than we are at present able to provide, but see Wenger who warns the analyst against imbuing the community of practice with "a concreteness they do not actually possess" (Wenger 1998:61).

9. Hay et al. (2008) also employ local frequency counts in their analysis of frequency effects, for the same reasons i.e. local lexical items and placenames which are common in New Zealand appear with much higher frequency counts in the ONZE corpus than in the CELEX lexical database. For further discussion of frequency and TH-Fronting in this corpus, see Clark and Trousdale (2009).

10. This excludes the lexemes WITH, THINK and THING, and variants of THINK and THING which were not variable in this corpus and, in fact, seemed to be involved in a different sound change (see Clark and Trousdale 2009 for details of these exceptions).

11. See Tagliamonte 2006: 149-150 for details of how to compare the log likelihood of different runs of the varbrul analysis to find the best 'fit'.

12. The overlap between syllable position and word position was too great in the analysis. Both produced a significant result but none more significant then the other. It was therefore necessary to collapse these two factor groups together into a single factor group.

13. Although there have been attempts to explain frequency effects in the structuralist tradition (e.g. Hammond 1999; Zuraw 2003), these have been made possible by incorporating an awareness of the relationship between language structure and language use.

14. The exact nature of the relationship between entrenchment and frequency of use (particularly within a corpus) is perhaps less straightforward than originally suggested in Langacker (1987). See Schmid (2007) for further discussion of this problem.

15. Stuart-Smith, Timmins, and Tweedie (2007) suggest a similar pattern for the loss of the voiceless velar fricative in Glasgow i.e. certain speakers are unaware that this is a supra-local change affecting British English generally and instead regard the change as wholly Glaswegian.

16. See the discussion of 'bad Scots' in Aitken (1982).

17. Competing with the band is a rite of passage to full membership – until this is achieved, it is very difficult to claim membership of the band.

References

Adger, David, and Jennifer Smith
    2005   Variation and the Minimalist Program. In *Syntax and Variation: Reconciling the Biological and the Social*, Leonie Cornips, and Karen Corrigan (eds.), 149-178. Amsterdam: Benjamins.

Aitken, Adam J.
    1982   Bad Scots: Some Superstitions about Scots Speech. *Scottish Language* 1: 30-44

Anderson, Bridget
    2002   Dialect levelling and *ai* monophthongisation among African American Detroiters. *Journal of Sociolinguistics* 6: 86-98.

Baayen, R. Harald, Richard Piepenbrock, and Leon Gulikers
    1995   The CELEX Lexical Database (Release 2) [CD-ROM]. Pennsylvania, PA: Linguistic Data Consortium, University of Pennsylvania.

Bayley, Robert
    2002   The quantitative paradigm. In *The Handbook of Language Variation and Change*, Jack K. Chambers, Peter Trudgill, and Natalie Schilling-Estes (eds.), 117-141. Oxford: Blackwell.

Britain, David
    2002   Space and spatial diffusion. In *The Handbook of Variation and Change*, Jack Chambers, Peter Trudgill, and Natalie Schilling-Estes (eds.), 603-637. Oxford: Blackwell.

Britain, David, and Kazuko Matsumoto
    2005   Language, communities, networks and practices. In *Clinical Sociolinguistics*, Martin J. Bell (ed.), 3-14. Oxford: Blackwell.

Borgatti, Stephen. P., Martin G. Everett, and Linton C. Freeman
    2002   *Ucinet for Windows: Software for Social network analysis.* Harvard: Analytic Technologies.

Bybee, Joan
    2001   *Phonology and Language Use.* Cambridge: Cambridge University Press.
    2007   *Frequency of Use and the Organization of Language.* Oxford: Oxford University Press.

Cameron, Deborah
    1997   Demythologizing sociolinguistics. In *Sociolinguistics: A Reader and Coursebook*, Nikolas Coupland, and Adam Jaworski (eds.), 55-68. London: Macmillan Press Ltd.

Clark, Lynn, and Trousdale, Graeme
    2009   The role of token frequency in phonological change: evidence from TH-Fronting in east-central Scotland. *English Language and Linguistics* 13 (1): 33-55.

Croft, William
  2009   Towards a social cognitive linguistics. In *New Directions in Cognitive Linguistics*, Vyvyan Evans, and Stéphanie Pourcel (eds.), 395-420. Amsterdam: Benjamins.
Croft, William, and D. Alan Cruse
  2004   *Cognitive Linguistics*. Cambridge: Cambridge University Press.
Docherty, Gerard
  2008   An evaluation of usage-based approaches to the modeling of socio-phonetic variability'. Paper presented at *17th Sociolinguistics Symposium (SS17)*, Amsterdam, the Netherlands.
Dyer, Judy
  2002   "We all speak the same round here": Dialect levelling in a Scottish-English community. *Journal of Sociolinguistics* 6 (1): 99-116.
Eckert, Penelope
  2000   *Linguistic Variation as Social Practice: The Linguistic Construction of Identity at Belten High*. Language in Society 27. Oxford: Blackwell.
  2002   Constructing meaning in sociolinguistic variation. Paper presented at the annual meeting of the American Anthropological Association, New Orleans, LA, November 2002. Available at http://www.stanford.edu/~eckert/AAA02.pdf
Foulkes, Paul, and Gerard Docherty
  1999a   Derby and Newcastle: instrumental phonetics and variationist studies. In *Urban Voices: Accent Studies in the British Isles,* Paul Foulkes, and Gerard Docherty (eds.), 1-24. London: Arnold.
  2006   The social life of phonetics and phonology. *Journal of Phonetics*, 409-438
Foulkes, Paul, and Gerard Docherty (eds.)
  1999b   *Urban Voices: Accent Studies in the British Isles*. London: Arnold.
Geeraerts, Dirk
  2001   On measuring lexical convergence. In *Linguagem e Cognição. A Perspectiva da Linguística Cognitiva*, Augusto Soares da Silva (ed.), 51-61. Braga: Associação Portuguesa de Linguística.
Guy, Gregory R.
  1988   Advanced varbrul analysis. In *Linguistic Change and Contact,* Kathleen Ferrara, Becky Brown, Keith Walters and John Baugh (eds.), 124-136. Austin, TX: Department of Linguistics, University of Texas at Austin.
Guy, Gregory R.
  1993   The quantitative analysis of linguistic variation. In *American dialect research*, Dennis Preston (ed.), 223-249. Amsterdam: Benjamins.

Hammond, M.

1999   Lexical frequency and rhythm. In *Functualism and formalism in linguistics: General Papers*, Michael Darnell, Edith Moravcsik, Fredrick Newmeyer, Michael Noonan, and Kathleen Wheatley (eds.), 329-358. Amsterdam: Benjamins.

Hay, Jennifer

2001   Lexical frequency in morphology: is everything relative? *Linguistics* 39 (6): 1041-1070.

Hay, Jennifer, Gregory R. Guy, Abbey Walker, and Benjamin Munson

2008   Coronal stop deletion revisited. Paper presented at University of Newcastle.

Hollmann, Willem, and Anna Siewierska

2007   A construction grammar account of possessive constructions in Lancashire dialect: some advantages and challenges. *English Language and Linguistics* 11 (2): 407-442.

Hudson, Richard

1996   *Sociolinguistics* (second ed). Cambridge: Cambridge University Press.

2007a   *Language Networks: The New Word Grammar*. Oxford: Oxford University Press.

2007b   English dialect syntax in Word Grammar. *English Language and Linguistics* 11 (2): 383-405.

Kemmer, Suzanne, and Michael Barlow

2000   A usage based conception of language. In *Usage-Based Models of Language*, Suzanne Kemmer, and Michael Barlow (eds.), 7-28. California: CSLI Publications

Kerswill, Paul

2003   Dialect levelling and geographical diffusion in British English. In *Social Dialectology. In Honour of Peter Trudgill,* David Britain, and Jenny Cheshire (eds.), 223-243. Amsterdam: Benjamins.

Kristiansen, Gitte

2003   How to do things with allophones: Linguistic stereotypes as cognitive reference points in social cognition. In *Cognitive Models in Language and Thought*, CLR 24, René Dirven, Roslyn M. Frank, and Martin Pütz (eds.), 69-120. Berlin/New York: Mouton de Gruyter.

2006   Towards a Usage-Based Cognitive Phonology. *International Journal of English Studies* 6 (2): 107-140

Kristiansen, Gitte, and René Dirven (eds.)

2008   *Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems*. Berlin/New York: Mouton de Gruyter.

Labov, William
  1972   *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.
  2002   Driving Forces in Linguistic Change. Paper presented at the *International Conference on Korean Linguistics,* August. Seoul National University. Available at
          http://www.ling.upenn.edu/~wlabov/Papers/DFLC.htm
Langacker, Ronald
  1987   *Foundations of Cognitive Grammar, Vol 1. Theoretical Prerequisites.* Stanford, CA: Stanford University Press.
  1991   *Foundations of Cognitive Grammar, Vol 2. Descriptive Application*. Stanford, CA: Stanford University Press.
Le Page, Robert. B, and Andree Tabouret-Keller
  1985   *Acts of Identity: Creole-based Approaches to Language and Ethnicity*. Cambridge: Cambridge University Press.
Macafee, Caroline
  1983   *Varieties of English around the World: Glasgow*. Amsterdam: Benjamins.
Mather, James Y., and Hans-Henning Speitel
  1986   *The Linguistic Atlas of Scotland, Volume III: Phonology.* London: Croom Helm.
Mathews, Tanya
  2005   Discourses of intergroup distinctiveness among adolescent girls. Ph.D. dissertation, Cornell University.
Mendoza-Denton, Norma
  1997   Chicana/Mexicana identity and linguistic variation: an ethnographic and sociolinguistic study of gang affiliation in an urban high school. Ph.D. diss., Stanford University.
Milroy, Lesley, and Mathew Gordon
  2003   *Sociolinguistics: Method and Interpretation*. Language in Society 34. Oxford: Blackwell.
Mompeán-Gonzalez, Jose
  2004   Category overlap and neutralization: the importance of speakers' classifications in phonology. *Cognitive Linguistics* 15: 429-469.
Moore, Emma
  2003   Learning style and identity: a sociolinguistic analysis of a Bolton high school. Ph.D. diss. University of Manchester.
Paolillo, John C.
  2002   *Analyzing Linguistic Variation: Statistical Models and Methods*. Stanford: CSLI.

Robinson, Christine
  2005   Changes in the dialect of Livingston. *Language and Literature* 14
              (2): 181-193.
Rothbart, Myron, and Lewis, Scott
  1998   Inferring category attributes from exemplar attributes: geometric
              shapes and social categories. *Journal of Personality and Social Psy-
              chology* 55: 861-872.
Sankoff, David, Sali Tagliamonte, and Eric Smith
  2005   Goldvarb X: A Variable Rule Application for Macintosh and Win-
              dows. Department of Linguistics, University of Toronto.
              http://individual.utoronto.ca/tagliamonte/Goldvarb/GV_index.htm.
Schmid, Hans-Jörg
  2007   Does frequency in text really instantiate entrenchment in the cogni-
              tive system? Paper presented at the Theme session on Corpus Driven
              Cognitive Semantics at the 10th International Cognitive Linguistics
              Conference, Jageiellonian University, Poland, July.
Stuart-Smith, Jane, and Fiona Tweedie
  2000   Accent Change in Glaswegian: A Sociophonetic Investigation (Final
              report to the Leverhulme Trust). (F/179/AX).
Stuart-Smith, Jane, and Claire Timmins
  2006   The role of the lexicon in TH-fronting in Glaswegian. *The Power of
              Words: Essays in Lexicography, Lexicology and Semantics*, Coste-
              rus 163, Graham D. Caie, Carole Hough, and Irené Wotherspoon
              (eds.), 171-184. Amsterdam: Editions Rodopi.
Stuart-Smith, Jane, Claire Timmins, and Fiona Tweedie
  2007   "Talkin' Jockney?" Variation and change in Glaswegian accent.
              Journal of Sociolinguistics 11 (2): 221-260.
Tagliamonte, Sali
  2006   *Analysing Sociolinguistic Variation*. Cambridge: Cambridge Univer-
              sity Press.
Taylor, John
  2002   *Cognitive Grammar*. Oxford: Oxford University Press.
Tummers, José, Kris Heylen, and Dirk Geeraerts
  2005   Usage-based approaches in Cognitive Linguistics: a technical state
              of the art. *Corpus Linguistics and Linguistic Theory* 1: 225-261.
Tollfree, Laura
  1999   South East London English: discrete versus continuous modelling of
              consonantal reduction. In *Urban Voices: Accent Studies in the Brit-
              ish Isles*, Paul Foulkes, and Gerard Docherty (eds.), 163-184. Lon-
              don: Arnold.

Trousdale, Graeme
  2002   Variable ambisyllabicity. *English Language and Linguistics* 6: 267-
              282.
Watson, Kevin
  2006   Language & gender meets theoretical phonology: bridging the peda-
              gogic gap? Paper presented at the *BAAL Special Interest Group on
              Issues in teaching Language, Gender and Sexuality in Further and
              Higher Education*. Liverpool Hope University.
Wells, John C.
  1982   *Accents of English: the British Isles. Vol. 2* Cambridge: Cambridge
              University Press.
Wenger, Etienne
  1998   *Communities of Practice: Learning, Meaning, and Identity*. Cam-
              bridge: Cambridge University Press.
Williams, Anne, and Paul Kerswill
  1999   Dialect levelling: change and continuity in Milton Keynes, Reading
              and Hull. In *Urban Voices: Accent Studies in the British Isles*, Paul
              Foulkes, and Gerard Docherty (eds.), 141-162. London: Arnold.
Wolfram, Walt, and Natalie Schilling-Estes
  1996   Dialect change and maintenance in a post-insular community. In
              Focus on the USA, Edgar Schneider (ed.), 103-148. Amsterdam:
              Benjamins.
Zuraw, Kie
  2003   Probability in language change. In *Probabilistic Linguistics*, Rens
              Bod, Jennifer Hay, and Stephanie Jannedy (eds.), 139-176. Cam-
              bridge MA: MIT Press.

Overview

As a schematic index to the volume, the following chart positions the chapters on three thematic dimensions: the language of description, the specific linguistic variable under investigation (and the overal linguistic field to which it belongs), and the type of variation being studied. (Diatopic variation involves variation in space, like geographically distinct dialects or national varieties. Diastratic variation involves sociostylistisc factors. Diachronic variation involves variation in time. The classification is not meant as categorial; it only specifies the dominant perspective of the chapters.)

			diatopic	diastratic	diachronic
		LEXICON			
Geeraerts and Speelman	Dutch	the lexical field of the human body	x		
Soares da Silva	Portuguese	the lexical fields of football and clothing	x		x
Robinson	English	the lexical item *awesome*		x	
Peirsman et al.	Dutch	the lexical field of religion			x
		MORPHOLOGY AND SYNTAX			
Szmrecsanyi	English	the genitive construction	x	x	
De Vogelaer	Dutch	nominal gender	x		
Colleman	Dutch	the ditransitive construction			x
		PRONUNCIATION			
Kristiansen	Spanish	accent recognition	x	x	
Berthele	German	accent evaluation	x		
Clark and Trousdale	English	th-fronting			x