
Corpus-linguistic applications
Current studies, new directions
Edited by
Stefan Th. Gries, Stefanie Wulff
and Mark Davies

Corpus-linguistic applications

Current studies, new directions

LANGUAGE AND COMPUTERS: STUDIES IN PRACTICAL LINGUISTICS

No 71

edited by
Christian Mair
Charles F. Meyer
Nelleke Oostdijk

Corpus-linguistic applications

Current studies, new directions

Edited by
Stefan Th. Gries
Stefanie Wulff
Mark Davies

Rodopi

Amsterdam - New York, NY 2010

Cover image: Morguefile.com

Cover design: Pier Post

The paper on which this book is printed meets the requirements of "ISO 9706:1994, Information and documentation - Paper for documents - Requirements for permanence".

ISBN: 978-90-420-2800-5
E-Book ISBN: 978-90-420-2801-2

Printed in The Netherlands

Contents

Introduction

Stefanie Wulff

University of North Texas, Denton

Stefan Th. Gries

University of California, Santa Barbara

Mark Davies

Brigham Young University

The eighth conference of the American Association for Corpus Linguistics was hosted by the Brigham Young University in Provo, Utah, from 13–15 March 2008. Over the course of three days, more than 120 papers were presented, most of them at Brigham Young's own Aspen Grove conference center, located right at the bottom of the beautiful Mount Timpanogos. Next to this unique scenery, the conference organizers indulged the conference participants with their welcoming attitude and meticulous organization throughout. On behalf of everybody, the two editors listed here first would like to compliment the conference organizers once again on making this a truly stimulating experience.

Roundabout a year later, this volume comprises fifteen papers that withstood the scrutiny of an intensive blind review and revision process. The editorial team, Stefan Th. Gries, Stefanie Wulff, and Mark Davies, tried their best to find at least two qualified reviewers for every submission. From the feedback that we received from the contributors, we have come to understand that many conceived of the review process as critical and intense, but also constructive, a feedback we are quite satisfied with. However, while each contribution to the volume as it stands clearly deserves to be published, we would like to emphasize that the reverse conclusion is not justified: papers that were submitted but did not make it through the review process are not necessarily less worthy of publication, not to speak of the overwhelming majority of conference presentations that were not submitted for publication in the present volume in the first place.

Instead, a major concern of the editorial team was to compile a volume that would provide a coherent and representative sampling of the conference presentations in terms of the topic areas and corpus-linguistics methodologies covered. In the ideal case, a brief look at the table of contents alone would give the reader a quick answer to the question that the title of this volume raises: what *are* the current studies and new directions? Or in other words: what's hot in corpus linguistics in 2008/2009? As we see it, the papers fall into four quite coherent categories, and so we divided this volume into four sections: a

diachronic section, a *genre* section, a *forms of functions* section, and a *methods and tools* section. Needless to say, most papers can arguably belong to more than one category, or even another category not listed here; accordingly, the purpose of this grouping is less to bin the papers, but instead to highlight larger trends that become obvious when looking at the selected papers in this volume, as well as the conference presentations *in toto*. In the following, we briefly describe these four major sections and the contributions therein.

In many ways, corpus linguistics was born in an attempt to adequately describe and analyze contemporary language use based on naturally occurring language data as opposed to prescriptive grammar books. In 2008, a major emerging trend in the field is to use established corpus-linguistic tools and apply them to *diachronic* data. An increasing number of diachronic corpora are becoming publicly available, such as Mark Davies' *Corpus Del Español* (CDE), which comprises 100 million words from the 1200s to the 1900s; this apparently stimulates historical research, particularly on Spanish and Portguese (three out of the four papers in the *diachronic section* of this volume are either language). That corpus linguistics is no longer in its infancy becomes obvious when we consider the range of linguistic phenomena the development of which the authors examine here: the section includes studies of morphological, lexical, and overall content development.

The widest focus with regard to the time span taken into account is Miglio's contribution, in which she tracks the development of the Spanish adverb *dizque* from the 13th to the 20th century. Bringing together evidence from various corpora such as the afore-mentioned CDE and Real Academia's *Corpus diacrónico del español* (CORDE)/*Corpus de referencia del español actual* (CREA), Miglio shows that the diachronic development of *dizque* did not follow a straightforward path, but fluctuated with regard to frequencies of meanings, not least also depending on the specific text type in which it was attested.

Medina Urrea combines a diachronic perspective with sophisticated statistical methods to track the development of affix sequences in Spanish from the 16th to the 20th century. The measure of affixality Medina Urrea lays out is a complex one that combines normalized square counts, entropy, and economy values. The resulting morphological profiles license various new hypotheses about the development of Spanish dialects, such as the emergence of Mexican Spanish before the 18th century. In addition, Medina Urrea calculated Euclidean distances between the different morphological profiles to study the coming into existence of Mexican Spanish as a distinct dialectal system and the development of Peninsular Spanish.

Rudanko focuses on recent distributional changes of infinitival and gerundial complements of the verb *submit* in English. A dramatic increase in the share of gerundial complements over the last 100 years seems to defy simple explanations in terms of the established meaning differences between the two complementation patterns. Instead, Rudanko argues that this "Great Complement Shift" could have been motivated by a complex interplay of the semantics of the

gerundial complementation pattern, *submit*'s preference for passives and passive-like lower predicates, and Undergoer/Patient-subjects as lower subjects.

Like Rudanko, Mota also tracks more recent diachronic developments. Her corpus samples of Portuguese journalese cover the period from 1991 to 1998. Applying Kilgarriff's (2001) corpus similarity measure, Mota finds that for texts on the same topic, there is a negative correlation between the time gap between their respective publication dates and their assessed similarity. Sufficiently large time gaps can even override thematic content such that two texts from sufficiently different time periods on the same topic are as dissimilar to each other as two texts on different topics. A comparison with similarity measures based on different kinds of frequency lists (comprising only upper and lower case words, respectively) yield highly comparable results.

Another main root of corpus linguistics is that of lexicography, and we are indebted to early lexicographic work (such as that by Sinclair, Krishnamurthy, and others) for some of the most valuable and impressive contributions to the field. In analogy to more recent theorizing about language inside and outside of corpus linguistics (think Hunston and Francis' [2000] *Pattern Grammar*, Biber et al.'s [1999] *lexical bundles*, or Goldberg's [1995, 2006] *Construction Grammar*, for example), the *form of functions* section in the present volume illustrates that the scope of corpus linguistics has widened considerably ever since. The focus of corpus linguistics is by far no longer restricted to the investigation of lexical phenomena: the three papers address questions above and beyond the word level, including tags, semantic prosody, and grammatical subjects.

Columbus compares the meaning and functions of four invariant tags (*eh*, *yeah*, *no*, and *na*) in New Zealand, British, and Indian English. She outlines the most pertinent meanings and their functions and draws attention to potential sites of cultural misunderstandings as well as implications thereof for the adequate development of TESOL materials.

Dilts presents an attempt to systematically assess the semantic orientation of nouns in the *British National Corpus* by means of looking at the semantic coloring of the adjectives they co-occur with. These measures of the nouns' semantic orientation are then compared with semantic preference ratings for the category 'pleasure' from ANEW (Bradley and Lang 1999). The results suggest that there is indeed a stable correlation between semantic orientation and semantic preference. More specifically, negatively oriented nouns generally prefer to collocate with positive adjectives, while in the reverse case, only few positively oriented nouns collocate with negative adjectives.

Zdorenko enriches the *forms of functions* section by looking at a grammatical phenomenon, subject omission, in Russian. She finds that subject omission is highly register-dependent, with omission being most frequent in informal, spontaneous conversations (thereby going beyond previous studies, which focused exclusively on written language alone). Similarly, subject omission is most frequent with first and second person subjects, and it is intricately tied to specific verbs. In combination, these results point to the

ongoing grammaticalization of certain high-frequency morpho-lexical combinations into discourse markers.

A third dimension at which contemporary corpus linguistics has clearly evolved from its inceptive stage concerns the recognition of corpora not as unique gestalts, but complex compilations of various sub-corpora from different speakers in different genres and registers (not to mention dialectal and diachronic variation parameters). Accordingly, the *genre* section represents the various contributions to AACL 2008 that dealt with the potential impact of genre and register on linguistic variation.

Pho presents a contrastive analysis of the move structure and prominent linguistic features of research article abstracts and introductions in two different academic disciplines, educational technology and applied linguistics. She finds that overall, move structures and linguistic make-up vary more as a function of section than academic discipline. Furthermore, her analysis illustrates that moves are best understood as complex clusters of morphological, lexical, and syntactic features as opposed to individual markers.

Csomay and Cortes also raise our awareness of the intra-textual dependency of meaning and function. They examine how the discourse functions of lexical bundles in classroom discourse vary as a function of when/where in the course of a classroom session they occur. Bundles expressing stance give way to referential and discourse-organizing bundles, the number of which increases as the discourse unfolds.

Diniz also zooms in on university discourse. Using data from the *Michigan Corpus of Academic English*, she shows that professors use various modal verbs (including *should* and *could*) as well as lexical verbs (such as *recommend* and *suggest*) as markers of indirect, polite orders as opposed to straight-out directives. She illustrates this communicative function by extracting the most pertinent (semi-)fixed phrases found in her data.

Fitzpatrick and Bachenko make a valuable contribution to the *genre section* as they look into a genre decidedly outside of academia. On the basis of a self-compiled "forensic" corpus of criminal statements, police interrogations, and testimonies, they set out to examine the validity of various linguistic cues that have been identified in previous research to indicate insincerity on the part of the speaker, such as hedges, negative forms, or noun phrase changes. The near 75% prediction accuracy they yield, in spite of some limitations regarding the quality of their corpus data, suggests that their unique approach to deception detection is a promising avenue for future research.

Last but not least, maybe the most dramatic changes that the field of corpus linguistics is witnessing these days concerns its methodologies. Long gone are the days when all corpus linguists had at their disposal was a concordance, a frequency list, and maybe a list of collocations. As evidenced in the *methodology and tools* section, the field of corpus linguistics is rapidly being enriched with methodological expertise borrowed from other fields such as statistics, computational linguistics, and even artificial intelligence.

Gries is concerned with the crucial role that dispersion and adjusted frequencies should play in corpus-linguistic analyses. From a comprehensive comparison of various measures of dispersion and adjusted frequencies, paired with a comparison with psycholinguistic data, a complex picture emerges in which no single measure reveals itself as an "all purpose" solution. Instead, Gries cautions us to be aware of the limitations and advantages of the various measures, and to decide in favor of one after systematic comparison on a case-by-case basis.

Another contribution that sheds new light on a persistent issue in corpus-linguistic methodology is Cox's discussion of the functionality of probabilistic part-of-speech-tagging when applied to a corpus of Mennonite Low German as an example of a comparatively small minority language corpus. His findings suggest that various interrelated factors, including the size of the training sample, the complexity of the tag set, and issues of orthographic normalization impact tagging accuracy. What is more, the impact of these factors shifts during the training phase of the tagger.

Teich and Fankhauser present a multivariate approach to the question what characterizes the meta-register of scientific writing, and how to best assess the similarity of the sub-corpora comprising that meta-register. By pooling various statistical and data mining techniques, including feature ranking, clustering, and classification, Teich and Fankhauser identify various distinctive features of the meta-register of their scientific writing corpus in contrast to other kinds of writing, and also identify features associated with discipline-specific forms of writing.

Finally, Bloom et al. provide a nice example of how corpus linguistics can have direct applications outside the field of linguistics, specifically in marketing research. They present a grammatically motivated system for the automated learning of appraisal extraction patterns from data such as product reviews and movie reviews. Their automated system performs comparable to manual extraction.

In sum, the papers in this volume, as different as they are, convey at least one message quite univocally: corpus linguistics grows and prospers once we dare to problematize seemingly innocent notions and assumptions such as the following:

- corpus representativeness: different corpora vary by register and genre, and also within as a function of intra-textual structure;
- the word as the central unit of investigation: widening the scope of corpus-linguistics above and beyond the word level does justice to contemporary theories of language and grammar;
- the validity of frequency counts and other established methods: we are only beginning to understand the complex ways in which frequency effects pertain to questions of grammar and language, and accordingly, we need to remain open to new and alternative ways of counting, measuring, and weighing frequency information obtained from corpus data;

- monofactorial explanations: several studies in this volume demonstrate how multifactorial analyses help us understand the complex nature of linguistic processes better than monofactorial studies have done in the past.

These recognitions have immediate consequences for future directions in the field. Corpus linguistics is moving away from analyses based on a limited set of examples of naturally occurring language extracted with the help of ready-made software packages. Instead, corpus linguists are increasingly sophisticated in compiling, extracting, and evaluating their data in complex and innovative ways. Today's wordsmiths need to be experts in all these domains.

References

Biber, D., S. Johansson, G. Leech, S. Conrad, and E. Finegan (1999), *Longman grammar of spoken and written English*. London: Longman.

Bradley, M., and P. Lang (1999), *Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings*. Technical report C-1, Gainesville, FL. The Center for Research in Psychophysiology, University of Florida.

Goldberg, A. E. (1995), *Constructions: a construction grammar approach to argument structure*. Chicago: The University of Chicago Press.

Goldberg, A. E. (2006), *Constructions at work: the nature of generalization in language*. Oxford: Oxford University Press.

Hunston, S., and G. Francis (2000), *Pattern Grammar: a corpus-driven approach to the lexical grammar of English*. Amsterdam/Philadelphia: John Benjamins.

Online databases and language change: the case of Spanish *dizque*

*Viola G. Miglio**

University of California, Santa Barbara

Abstract

This paper explores the semantics and pragmatic usage of dizque*, an adverb used as an evidential strategy in Latin American Spanish (LAS), and charts its development from the 12th to the 20th century, concentrating on changes during the Colonial period, comparing data from online databases such as Mark Davies's Corpus del Español (CDE) and the Real Academia's Corpus diacrónico del español (CORDE)/Corpus de referencia del español actual (CREA) to a printed collection of Colonial texts (Company Company, 1994). The paper shows that* dizque *emerges very early as an impersonal form, which can be construed as an evidential strategy (13th century), and declines from the 17th century onwards according to CDE and CORDE, only to be found again in the 20th century with a different distribution. I also compare the use of* dizque *to mark disbelief with its later meaning as an evidential marker (17th century onwards), and show that fluctuations in the use of the form are related both to the evolution of its meaning and to the text type in which it appears. Despite some limitations due to the size or type of documents sampled, I conclude that online databases such as CDE, CORDE, and CREA are invaluable tools to establish trends in language change and to understand diatopic variation.*

1. *Dizque*: an introduction

In modern Spanish, mostly in Latin American usage, there exists a form, *dizque*, which is by and large used as a sentential adverb meaning 'allegedly, supposedly.' Even in a simple statement such as the preceding one, there are a number of points that need to be clarified. What is the form's origin, for instance? Can we locate this form along a grammaticalization cline? Is it one word or two? *Dizque* exists both as a solidified adverbial form *dizque*, and as a form that preserves some traces of its verbal nature (*diz* < *decir* 'to say'): have the two forms evolved differently? It is necessary to tease apart the most commonly found functions, evidential and epistemic, but other uses, such as the mirative, are restricted to certain dialects. Although the form is mostly used as a sentential marker, it has also developed an adjectival meaning to mark a speaker's stance in relation to alleged or supposed information. The present paper addresses some of the chronological, semantic, and pragmatic aspects of *dizque*'s evolution, as well as considers the question of its diatopic distribution in modern Spanish.

The meaning of this form as evidential, intended strictly speaking as secondhand information (Aikhenvald, 2004; see below), can be observed in the following sentence from Galician author Fernández Flórez.

(1) Comí un plato de carne asada, con una cosa que **diz que** le llaman batatas […] (Fernández Flórez, *Volvoreta*, CORDE)
'I had a dish of grilled meat, with a side dish that *dizque* they call sweet potatoes.'[1]

In this case, the narrator reports that this is secondhand information, having no reason to doubt its veracity. The epistemic meaning, on the other hand, can be seen in the following example.

(2) […] se le cerraban los ojos, se despidió para irse, **diz que** a cenar, o más bien, creo yo, a dormir. (Ayala, *El fondo del vaso*, 1962, CORDE)
'His eyelids drooping, he left saying he was *dizque* going to dinner, but I rather believe he was going to sleep.'

Here, *dizque* indicates that the speaker does not believe the information, not that it is second hand. The form is clearly disambiguated by the follow-up adversative interjection *más bien, creo yo* 'I rather believe.' The majority of corpora used for this study are written corpora, and it is hard to find examples as unambiguous as the one above, although for speakers of Spanish varieties with productive *dizque*, intuitions as to its meaning are very clear, even if usage may vary from one dialectal variety to another.

In the present paper, I present a study of what I call "an evidentiality strategy" (following Aikhenvald, 2004) in Spanish, using Mark Davies's *Corpus del Español* (CDE), the Real Academia's CORDE, and CREA, as well as a printed collection of Colonial Spanish texts (Company Company, 1994, now also online at <http://www.iling.unam.mx/chem/>).

Specifically, I explore the following questions: Given that it is well known that *dizque* is commonly used in Latin America but not in Spain (see, for instance, Magaña, 2005), who says that the form is "practically obsolete" in Spain? Does this impressionistic evaluation of the usage of *dizque* withstand the scrutiny of the data to be found in the different online corpora? I also evaluate whether the CORDE/CREA present a different picture from the CDE and from the printed collection of Colonial texts examined here (Company Company, 1994). Another question concerns the differentiation between the two most common functions of *dizque* as an evidential marker and as an epistemic marker. I sketch the order in which they developed and offer a chronology of their evolution. Another aspect examined here is the potential influence of genre and register on the contemporary usage of *dizque*, as well as on the evolution of this form (see also López-Izquierdo and Miglio, in progress).

2. Comparing the corpora

This introduction is by no means meant to be exhaustive, but it will serve to familiarize the reader with the features of the corpora that have had some bearing on my research on *dizque*.

Mark Davies's *Corpus del Español* (henceforth CDE; <www.corpusdelespanol.org>)[2] comprises more than 100 million words from approximately 20,000 documents spanning from the 1200s to the 1900s. Of these 100 million words, approximately five million are from oral documents, many of which are of an academic or legal nature, or political interviews. Approximately one million words of the oral component are composed of data from the *Corpus oral de referencia de la lengua española contemporánea* (online at <http://www.lllf.uam.es/~fmarcos/informes/corpus/corpulee.html>). This section includes, therefore, some informal, conversational material, and it is restricted also in that it is comprised only of Peninsular Spanish texts.

The CDE can be queried in different ways by word or lemma, as well as by period and (for the 1900s) by genre (oral, fiction, news, or academic). Author- or country-specific queries are not possible, but this information can be obtained manually. Unlike the corpora of the *Real Academia de la Lengua* (RAE), the CDE also provides the frequency distribution of the form under consideration as a function of occurrence for every million words throughout the centuries. In the case of *dizque*, one sees clearly that the form is not very common overall (Table 1). Even just by observing the frequency of distribution of the forms, especially the decline of the *diz que* form and the relative increase of *dizque*, one can surmise that it has undergone a process of grammaticalization, solidifying from its original verb + complementizer collocation into an adverbial form (López-Izquierdo, 2005).

Table 1. Frequency distribution (occurrence per million words) of the *dizque* form throughout the centuries according to CDE.

Form	**1200s**	**1300s**	**1400s**	**1500s**	**1600s**	**1700s**	**1800s**	**1900s**
Diz que	121.0	22.6	32.8	6.9	9.2	4.2	6.0	0.6
Dizque	0	0	0.4	0.9	0.9	0.8	0.4	1.4

The *Real Academia de la Lengua* (RAE) has two online corpora, the Corpus Diacrónico del Español (CORDE), which is used for historical research, and the Corpus del Español Actual (CREA), which consists of recent texts spanning approximately the last 30 years of the language. CORDE contains approximately 250 million words from the first medieval texts to 1975. In both corpora, searches can be restricted in a variety of ways. Documents can be culled from different types of texts such as fiction, politics, economics, commerce and finance, historical prose, or narrative prose. A search can also be restricted by country, year (or any different period established by the user), author, or work. Unlike the CDE, the RAE and CORDE corpora do not provide statistical

information on how many words make up a specific period of the corpus or how frequent the form under scrutiny occurs in the period examined.

CREA contains about 160 million words from documents spanning a period of a little more than 30 years (from 1975 to 2004). Like CORDE, CREA is mostly a written corpus, although 10% (approximately 10 million words[3]) are from oral interviews or from oral corpora that were available as transcribed documents. These mostly come from Spanish institutions (see, for instance, various university corpora, such as the one from Compostela, Mérida, Alicante, and Alcalá de Henares). Some of the oral documents come from Latin America, and according to the corpus description, they make up 50% of the oral forms.

Documentos Lingüísticos de la Nueva España (DLNE), edited by Concepción Company Company in 1994, is a corpus of approximately 270,000 words that I analyzed in its hard copy form (730 pages; it has since become available in digital format online). DNLE comprises a number of different documents from Mexico spanning the Colonial period, from 1525 to 1816. The documents are limited to those originating in the Central Highlands of Mexico—Mexico City, Puebla, and the nearby states of Tlaxcala, Querétaro, Guanajuato, and Morelos. As for the type of document, Company Company (1994: 6) states that she intended the corpus to represent the colloquial speech of the Colonial period. The bulk of texts are letters, official complaints, wills, inventories, petitions, witnesses' reports at trials, and official reports. They were mostly written by Spanish emigrants to their families or by missionaries, which Company Company considers particularly important given missionaries' role in the "castilianization" of the indigenous population.

3. Contrasting the occurrence of *dizque* in different corpora

Comparing the evolution of the structure in different corpora (see Figure 1), as well as the frequencies from CDE collected in Table 1 above, we see that *dizque* is not a very frequent form in contemporary Spanish (columns in Figure 1 span from the 1200s, the darkest column, to the 1900s, the lightest column, as in Table 1).

A word of caution about the *diz que* form is in order. This form could in fact be a personal form "s/he says" and therefore differ in meaning from the evidential or epistemic strategy considered for this paper. It is, however, unlikely for *diz que* to be an apocopated personal form from the 14th century onwards, except in some dialectal varieties. In the data above, the most problematic cases—where there is some doubt as to the form's impersonal or personal meaning—are the 1200s, the darkest column.

This finding certainly clashes with the impressionistic view of speakers from, say, Mexico or Colombia, who state that *dizque* is used quite commonly.[4] On the other hand, Lapesa (1984: 592–93) has an interesting comment about the form, which is cataloged among the syntactic periphrases of Latin American usage:

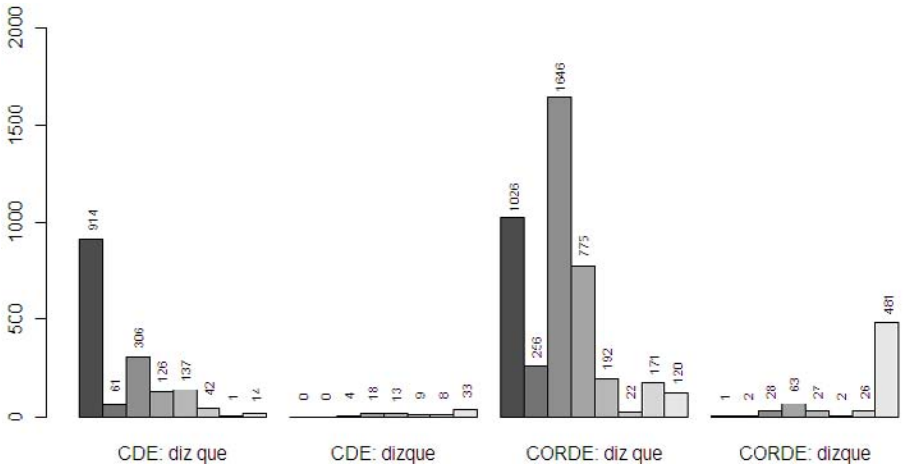


Figure 1. Evolution of *diz que* and *dizque* in CDE and CORDE.

(3) La antigua expresión impersonal **diz que,** indicadora de que el hablante repite noticias, rumores, tradiciones, etc. de origen impreciso, sobrevive en las formas *dizque, desque, isque, es que, y que,* no desconocidas, pero menos frecuentes en España.
'The older Spanish impersonal form *dizque,* used when the speaker repeats news, rumors, traditions, etc., of uncertain origin, survives in the forms *dizque, desque, isque, es que,* y *que,* which are not unknown, but less frequent in Spain.'

However, if one compares Iberian and Mexican Spanish through CORDE (Figure 2, which spans the 1500s, the darkest column, to the 1900s, the lightest column), *dizque* seems to be used equally as rarely in both countries in the 20th century: there are 47 instance of *dizque* in the Spanish texts, 35 in the Mexican texts, and a total of 263 instances for all Latin American texts.

A closer look reveals that the Spanish CORDE data comprise texts by Menéndez Pidal and Amado Alonso, who quote American Spanish examples. Other examples of Spanish *dizque* are from Galician authors such as Wenceslao Fernández Flórez, reflecting modern Galician use of a form *dizque* equivalent to that found in Latin American Spanish; from single authors whose provenance is uncertain but who often use the form (7 of those 47 forms are from Boronat y Barranchina); or from traditional, regional literary genres, such as Cantabrian tales. One has to agree with Lapesa that the form *dizque* is not very common in Spain.

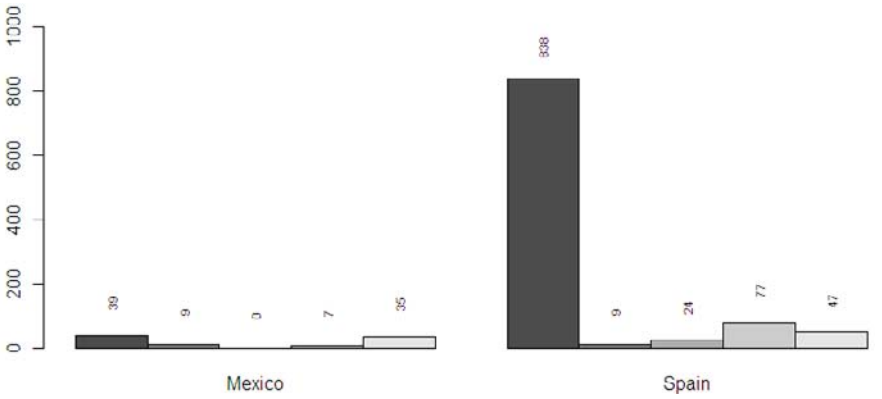


Figure 2. Comparison between Mexico and Spain through CORDE documents.

These figures indicate that standard Iberian Spanish rarely uses the form except where dialectal influence is to be noticed or because of a specific author's preference. This does not mean that there is no evidential strategy in Iberian Spanish, but rather that the need for evidentiality, or to evaluate the truth of reported information, is fulfilled by other forms (for instance, *dicen* 'they say' or the impersonal *se dice* 'it is said' form).[5] What needs to be explained, rather, is why there are so few cases in CORDE for Mexican Spanish. In this case, we can surmise that CORDE collects only formal prose, where *dizque* would not be found. As we will see below, from the 17th century onwards, *dizque* migrates to other registers, most probably to oral or informal prose, which are not abundantly represented in CORDE or CDE. If one adds the CREA cases, where contemporary fiction or narrative prose are included, another 265 cases (four of which are oral) are found, increasing the total number of hits for the 20th century to 485. *Dizque* is most commonly found in Colombia, Mexico, and the Dominican Republic, clearly showing that there is diatopic variation in usage, as well as a possibly diastratic distinction as to which register most readily allows the use of *dizque*.

4. Evolution of the form

4.1 The origins

Referring back to Figure 1 in the previous section, it could be said that the figure is somewhat misleading because of the high frequency of the form in the 13th century, especially in the prose from Alfonso X's scriptoria (historical prose). Those numbers include the apocopated personal form *diz que* 's/he/it says that'.[6] Although 96% of occurrences seem formulaic or impersonal in nature, one must consider the possibility that the high frequency of the *diz que* form here is simply

due to the presence of the third person present indicative of the verb *decir* 'to say' in its personal form.

For instance, we find examples where the form is likely to be personal because the subject is easy to recover:

(4) El abbat de Sant Andres de Spinareda por si e por so conuento se me enujo querellar e **diz que** quando quemara la su eglesia que quemaran hy los priuilegios que auie el monsterio e pediome por merced que yo mandasse saber la uerdat de quales priuillegios e de quales usos ouuieran [...] (CDE, Documentos castellanos de Alfonso X, 1200s)
'The Abbot at Saint Andrew's [...] on his behalf and on behalf of his convent sent a complaint **saying** that when the church burnt down, the charts of rights and possessions of the monastery were also burnt, and requested that I send him the information about which possessions and use of resources they had a right to [...]'

Because of the character of Alfonsine prose, which often portray geographical, historical, and mythical descriptions, the grammatical subject is not always so clearly recoverable, and in those instances where one can surmise that there is a grammatical subject, in fact, it is frequently a semantically bleached or atypical subject. It is often inanimate, as can be found in sentences such as "the document, the law, etc., says that [...]." Such sentences in modern Spanish would be impersonal or passive, the former subject cast as a locative of the type "in the law it is said" or as a similar construction. The following example is a particularly ambiguous one among the personal cases (the extended context in Spanish is provided below; only relevant parts are translated into English):

(5) E porende fueles menester que les veniese luz & claridat que les alunbrasen enla tenjebla enque estauan & desto auemos figura enla vieja ley / E **dize que** Jacob & el angel que lucharon toda la noche & quando la manñana vino dixo el angel a jacob dexa me ya que viene el alua & dixo Jacob sepas que non te dexare fasta que me des la tu bendiçion / Luego el angel bendixol en aquel lugar / E este jacob tomando lo spiritual mente podemos entender el humanal linaje / E la lucha que fezieron **diz que** fue de noche que se entiende la esperança que aujan quando vernjan la luz del alua & la luz del dia. (CDE, Castigos y documentos de Sancho IV, 1200s)
'[...] and of this we have an example in the old law / And **it says that** Jacob and the angel that they fought the whole night [...] / And the fight that they fought **says that** it took place at night [...]'

In this case, the first "it says that" seems to be personal, as it bears the personal markings without apocope of the final vowel, but the subject is not clear (and the complementizer *que* is repeated after "Jacob and the angel"). We can construe the subject to be the law, since in the previous sentence we find what I have loosely translated as "and of this we have an example in the old law."

However, the second instance of the *decir* verb *diz que* is so far from even the previous inference that it can only be interpreted as 'they say that, it is said that' (i.e., an impersonal form).

Even clearer is the example that follows, where *diz que* seems to repeat the Latin formula *dicitur* used when quoting the scriptures and corresponding to 'it is said (that).'

(6) **Unde dicitur**: At illi [continuo], relictis retibus (et navi) secuti sunt eum. * En ribera daquel mar farto Jhesu Cristo .v. mil omnes de .v. panes e .ii. peces. * En esta ribera a un castiello **que dizen** Corazaym, o **diz que** sera nodrido el Antechristo e(n) el enganador del sieglo. (CORDE, Alerich, Fazienda de Ultra Mar, ca. 1200)
'Where **it is said**: [...] On this shore there is a castle that **they say** (i.e., they call) Corazaym, where **says that** that the Antichrist will be raised [...]'

Thus, it is in the abundant use of the Alfonsine prose that the collocation *diz que* starts acquiring impersonal overtones. From a perusal of the forms in CORDE, 914 out of 952 tokens (96%) could be construed as impersonal in the prose of the 1200s: *diz que* was used here to corroborate facts that the author did not experience directly because they occurred in the ancient past, or it could apply to inference or hearsay. The predominance of impersonal-type forms is due to various factors: because of the context in which it was used, in historical or legal prose, even cases of personal occurrences of *decir* tended to have an inanimate subject, often 'the document' or 'the law.' The low agentivity of subjects in these cases promotes quasi-impersonal interpretations, which in turn are associated, little by little, to that specific collocation of *decir que*.

The genre itself is conducive to the need to express corroboration for facts that would certainly not be considered as such in modern historiography because they happened in the ancient past, are mythical in nature, and in any case do not come from trustworthy sources. However, the medieval author at this point never seems to question the veracity of the source; he merely wants to state for the record that he is reporting secondhand information: *[Atila]* ***diz que*** *fazie el cuemo el leon ferido […],* or '[Attila] *diz que* behaved like a wounded lion,' and talking about Roman times: *et* ***diz que*** *auia ally vna puente de canto con un arco muy grand que cogie este rio todo* or 'And *diz que* there was a stone bridge there whose arch spanned the whole river' (Alfonso X, Estoria de España, CDE, 1200s).

Another factor seems to be the use of this construction to translate formulas such as the synthetic passive *dicitur* 'it is said (that)' (at times also used as equivalent to *vocatur* 'it is called'). For these reasons, while many occurrences of *dizque* in texts from the 1200s are indeed personal, a closer reading of the examples from the corpora for this period reveals that 96% of the examples have characteristics that make them subject to a quasi-impersonal interpretation. In fact, M. López-Izquierdo (2005) argues that characteristics such as low agentivity

of subject and the apocopated form create the conditions permitting *diz que* to undergo a process of grammaticalization.

The apocopation of the form is a common phonological process in the 12th and 13th centuries affecting words ending in coronal sounds and a vowel, most often [–e] (Menéndez Pidal, 1987: 167–69). The forms *dize* and *diz*, along with other words with the same phonological make-up, coexist for a period, but usually the final *–e* is restored. The restoration of apocopated vowels in words such as *noch* for *noche* 'night' or *nuef* for *nueve* 'nine' is one of the often-cited characteristics of Alfonsine prose. As López-Izquierdo surmises in her 2005 paper, it is very likely that the survival of the apocopated form of *dize que* > *diz que/dizque* in the 13th century was due to a specialization of its meaning, connected to the impersonal, formulaic usage that we can appreciate from the quotations above. On the other hand, the sharp decline in usage of the form *diz* in the collocation *diz que* from the 14th century onward (CORDE reports 260 forms for that century and CDE 67) does not refer to the incipient impersonal usage, but to the personal form, which from that century onward is more likely to appear with the restored final vowel.

4.2 The first derived meaning: secondhand information and evidentiality

Although Nebrija (1492) does mention the form, some of the first grammarians of Spanish maintain that the form *diz que* is a shortened version of *dicen que* 'they say that,' (i.e., a plural form). This is what Juan de Valdés (quoted in López-Izquierdo, 2005, from *Diálogo de la lengua*, 1526) says: "También dezimos **diz que** por dizen, y no parece mal […]," "We also say *diz que* for they say, and it is considered correct [...]." However, Covarrubias, in his 1611 *Tesoro de la lengua castellana o española* says of *dizque*, states: "Palabra aldeana que no se debe usar en corte. Vale tanto como dicen que," "Peasant word that must not be used at court. It is the equivalent of 'they say.'" Bartolomé Jiménez Patón, speaking about contemporary Spanish usage in his *Elocuencia española en arte* (1604–1621, in the CORDE database), lists *dizque* among the old-fashioned particles that should not be used in formal speech because they abound only on the tongue of "ignorant rustics." The Real Academia's *Diccionario de Autoridades* (1732), on the other hand, still has the meaning of *dizque* as equivalent to the plural form of the verb 'to say' (i.e., *dicen que* 'they say').

What is important in these quotations is that the form is obviously in use, despite being considered correct at the beginning of the 16th century, only to be stigmatized at court a century later. The two contrasting views of its appropriateness suggest a shift in the locus of *dizque* from a cultivated register to a more oral, informal, and provincial usage in 17th-century Spain, which is the last of the centuries during the Colonial period (1500–1800), where we find it documented in writing with a certain frequency; from the 17th century to the 18th, CORDE examples drop from 200 to 22 for *diz que* and from 28 to 2 for *dizque*, and CDE examples go from 137 to 48 for *diz que* and from 13 to 9 for *dizque*.

The previous section illustrates that characteristics of the form *diz que,* such as a low agentivity subject and apocopation of final vowel, that erased

person information from the verb form, created an ambiguous interpretation, with one possible reading as an impersonal form with the meaning of 'it is said that.' As is typologically common, quotative markers usually evolve from forms of verbs meaning 'to say' (see examples in Aikhenvald, 1993: 4). In the 1300s, we see that *diz que* acquires the function of a quotative or hearsay marker. This is a common extension of an already existing grammatical category (verb of saying plus complementizer) to cover evidential-like meaning. The French conditional and the German subjunctive reflect a similar process of grammaticalization: the evidential meaning is not their sole or even primary meaning, which however extends in time to indicate source of information (Aikhenvald, 2004: 11).

Thus, *dizque* is a form that began as the collocation of a verb introducing indirect speech and its complementizer, and developed into an evidential strategy. It is not the only particle or collocation that indicates the speaker's stance to reality and information: in Colonial Spanish one finds many more, such as *tengo entendido que* 'I have heard that,' *sé decir y afirmo que* 'I can say and state that,' *soy informado (y muy certificado) que* 'I have been informed (and it has been guaranteed to me) that,' *tengo por cierto* 'I know for a fact,' *era fama que* 'it was commonly known that,' etc. (see Miglio, in press, for an analysis of these other forms).

In the 14th century, one can still find examples, such as (7) below, which may be considered ambiguous between a personal and impersonal reading (see Figures 3 and 4 for a tally of personal and impersonal instances). The author here reports on the death of Bishop Isidore of Seville in 636, and the sequence of the two forms of 'to say' in *diz que les dixo* 'it is said that he said to them' clearly disambiguates the form *diz que*. It cannot be Saint Isidore doing the saying in the first *diz*, since he is also the subject of the second *dixo*. Thus, it could be construed as an impersonal construction, almost adverbial in nature, meaning 'people say, the rumor goes.' There is, however, an antecedent in the previous line that could be the grammatical subject of *diz* (i.e., 'story/history'), which could be translated as 'so the story goes.' Despite the possibly personal subject, in the wider context of this sentence, it is very close to an impersonal form:

(7) Commo murio santo ysidro arçobispo de seujlla & fue sepultado enesa çibdat /. // Andados tres aços del Regnado del Rey sintilla **Cuenta la estoria** que santo ysidro despues que fue venjdo del Consejo que fuera fecho en toledo & predicando vn dia al pueblo **diz que** les dixo enaquella predicaçion muchas Cosas de profeias & de grant deuoçion & les dixo de su muerte que era çerca [...] (Pedro Afonso de Barcelos, tranlated by anonymous, Crónica de 1344 I, CDE)
'How Saint Isidore archbishop of Seville died and was buried in that city [...] **The story goes** that Saint Isidore, after the Council of Toledo, and preaching one day to the people **(the story) says** that/they say that he told them in that sermon [...] how his death was near [...].'

The following example is also culled from the same type of historical prose, where *dize que* and *diz que* cannot be interpreted as anything but *se dice que* 'it is said that':

(8) En el cviijo Capitulo **dize que** se quisieron alçar los de valencia contra abenjab por que conbidara un. dia a comer al [Ç]id & quisieran le matar ssi non por mjedo del [Ç]id. // En el cjx capitulo **diz que** abenjab sopo que los almoravides venjen a valencia e eran ya çerca de la villa e enbio dezir al [Ç]id quenon querie con el aver pleyto njnguno. (Juan Manuel, Crónica abreviada, 1319–1325, CDE)
'In chapter 108 **it says** [personal form?] that the people of Valencia wanted to revolt against Abenjab [...]. // In chapter 109 **it is said that** Abenjab found out that the Almoravids were coming to Valencia [...].'

As a hearsay marker, or marker of common knowledge, we find *diz que* in a number of formulaic expressions such as proverbs:

(9) Ia se yua ueyendo; Tebas en estrechura. Ca el Rey alexandre; dauales grant pressura. Mostraua les afirmes; que auia grant rancura. / Era mal quista; tebas de su frontera. Cuemo **diz que** mal debdo; a mal tiempo espera. Conteio a Tebas; dessa misma manera. (Anonymous, Libro de Alexandre, first half of the 13th century, CDE)
'Thebes had it coming, the city was in a pinch because King Alexander kept it under pressure [...] Thebes was not well liked by neighboring cities. As **they say,** a bad debt invites bad times (i.e., what comes around goes around), and this is exactly what happened to Thebes.'

In the example above there is really no specific subject, except *for* the speakers that use that proverb. As such, this can be considered a formulaic usage of the form.

At the end of the 15th and beginning of the 16th century, examples emerge that clearly show that the form is now considered a solidified adverbial and not a verb plus complementizer, as can be seen in Juan de Valdés's comment above (see López-Izquierdo, 2005, for a detailed evolution of the grammaticalization of the form, including the syntactic changes that affected it). In the example below from a legal document, one can appreciate that *que* 'that' is no longer a complementizer, since it is no longer at the beginning of the sentence, and the complementizer role is in fact fulfilled by a relative pronoun, which refers back to *paredes* 'walls.'

(10) Mandaron dar mandamiento para los alarifes que vean unas paredes de su mujer de Diego Gonçalez questan **diz que** armado sobrellas [...] de lo edificado por la Villa en la casa del alhondiga e lo vengan a declarar el primero ayuntamiento. (Anonymous, Libro de Acuerdos del Concejo Madrileño, 1498–1501, CORDE)

> 'They sent for the builders to check out some walls belonging to Diego González's wife, which **they say** are built on top of what was built by the Municipality at the public grain exchange house and that they should declare it to be the first Town Hall.'

In this example it is hard to construe *diz que* as anything but an adverbial meaning 'supposedly, rumor has it.' These meanings of secondhand information, hearsay, and common knowledge (*vox populi*)—what can be considered an evidential strategy (in the sense of Aikhenvald, 2004)—are the original ones that arise as the form of the verb *decir* 'to say' plus the complementizer *que* are grammaticalized into the adverbial form.

It should be noted that the adverbial form in its early usage here does not entail any mistrust of the source of information, but merely that the information is not directly obtained. There is no example in these early centuries of what can be called the epistemic meaning of *dizque*, with which a speaker using *dizque* might eschew responsibility for the information communicated, hint at the untrustworthiness of the information, or even imply that s/he believes it to be a lie. This is an important point that has been overlooked by other authors (for instance, Magaña, 2005).

In the case of Spanish and Romance languages in general, one should define the usage of tenses or adverbial particles such as *dizque* as evidential strategies, rather than evidentials proper. Following Aikhenvald's research, evidentiality is considered here as a grammatical category identifying the source of the information on which a statement is based (Aikhenvald, 2003: 1, 2004: 3). In many cultures it is of extreme importance to state clearly how one has acquired knowledge of something (directly witnessed or reported, through which sensory organs, inference etc.). Evidentiality is therefore obligatorily expressed in those languages. For this reason, *dizque* and the other markers of evidentiality in Romance languages should be termed evidentiality strategies, rather than evidentials proper, since they are not obligatorily expressed.

There are, however, many Amerindian languages in contact with Spanish (Quechua, Aymara, and Amazonian languages, or in Mexico, Chinantec, Mixtec, as well as some Uto-Aztecan and Mayan languages; Aikhenvald, 2004: 291) that do require the obligatory expression of some evidential category. This is likely to have reinforced the use of *dizque* as an evidential strategy in Latin American Spanish, even after its usage declined in Spain. This would be consistent with Aikhenvald's typological evidence showing that evidentiality easily spreads across different language families (ibid., 288ff). It is interesting to notice that *dizque* is, in fact, also found in Brazilian Portuguese (ibid., 179) and may well have had a parallel evolution as the Spanish form (i.e., common in Latin America), lost in the country where it originated. The difference is that Galician has also retained it in current usage (Victoria Vázquez, p.c.).

4.3 Epistemic meaning of *dizque*

Summarizing the previous section, *dizque* can be said not to be an evidential proper because the category is not obligatorily expressed in Spanish, but it is an evidentiality strategy because it allows Spanish to express an evidential meaning if need be. Its origins are to be found in the legal and historical prose of the 13th century, when the impersonal and formulaic usage of the verb *to say* and its complementizer become little by little welded together and their meaning changes into a quotative and hearsay marker (i.e., an evidential strategy).

In this section I would like to show that the epistemic meaning of *dizque*, or the meaning whereby the speaker expresses mistrust in the information s/he reports, should be considered secondary and derivative of the evidential one. Aikhenvald (2004) once again provides cross-linguistic evidence from the world's languages that the epistemic meaning is an extension of the role of reported evidentials (see pages 141, 179–182), and through CDE and CORDE there is chronological evidence that this is definitely the case for Spanish.

Around the beginning of the 1600s two different types of examples start appearing in the data along with the clearly evidential meanings exemplified in the previous section. These are examples that border on the absurd, or that at least display some clearly questionable information. At times, they betray the speaker's emotional involvement regarding the true value of the statement s/he is reporting, which the speaker clearly doubts (epistemic meaning), or show that the speaker is surprised by the statement s/he is making (mirative meaning).

The following example (11) is an early precursor of the epistemic meaning (from a *Carta de Relación* dated 1525), because it is such an absurd context that it is debatable whether the writer really did believe the information, but it is nevertheless impossible to be certain whether *dizque* repeated twice in close quarters here indicates disbelief.

(11) Y a los yndios que de aca yban con los christianos **diz que** guardaron para comer, y a los christianos hechaban en la laguna porque **diz que** los han provado y son duros y amarga la carne dellos. (Company Company 1994: 25)
'And the Indians that left here with the Christians they kept **they say** in order to eat them, but the Christians they would throw in the lagoon because **they say** that they have tried them and their flesh is hard and bitter.'

The following two examples are about 100 years older and clearly exemplify the epistemic meaning:

(12) Su madre **diz que** es donzella antes del parto y después; en aquesso no me meto, que verdad deve de ser. (José de Valdivieso, Romancero espiritual, 1638, CORDE)
'His mother **they say** is a virgin before and after his birth; I will not argue with that, as it must be true.'

And again the example below leaves no doubts as to an interpretation where the speaker contradicts the information he is reporting. Here *dizque* was not translated as 'they say,' because it clearly no longer corresponds to the evidential meaning, but to a particle that depicts a much more emotive involvement on the part of the speaker:

(13) […] **diz que** era pobre como yo soy abadesa. (José de Valdivieso, Romancero espiritual, 1638, CORDE)
'[…] *diz que* he was as poor as I am an abbess (i.e., he maintained he was poor, but I am obviously not convinced!).'

At the same time we find another type of meaning, still current in certain dialects of Latin American Spanish such as Colombian (see Travis, 2006, for this dialectal variety). This is the mirative meaning, in the sense of a grammatical element that introduces emotionally charged, surprising, or unexpected information.

(14) ¡Válgame las cuatro patas del caballo de Longino! ¿**Diz que** tengo de decir lo que no he visto, ni sé …? (Tirso de Molina, Quien da luego da dos veces, 1616, CDE)
'By the four legs of Longinus's horse! **Am I really supposed** to say what I did not see and do not know?'

It is the acquisition of these emotionally charged meanings, the epistemic and the mirative, that mark a change in the course of the distribution of the form. The questioning of the veracity of the reported information and surprise are often tinged with negative flair, so it becomes more and more difficult to use the adverbial as a simple evidential strategy. This can be appreciated in the jocular use Sor Juana Inés de a Cruz makes of the form in this poem:

(15)

Érase una Niña,/ como digo a usté,	There was a girl, I am telling you,
cuyos años eran, /ocho sobre diez.	Who was 18 years old,
Esperen, aguarden, /que yo lo diré.	Wait, be patient, now I will tell you.
Ésta (qué sé yo, /cómo pudo ser),	This girl (and I know what she was like)
dizque supo mucho, /aunque era era mujer.	**supposedly** knew a lot, even if she was a woman
Esperen, aguarden, /que yo lo diré.	Wait, be patient, now I will tell you.
Porque, como **dizque** /dice no sé quién,	Because, as I-don't-know-who **supposedly** says,
ellas sólo saben / hilar y coser ...	They [women] only know how to spin and sew ...

(del CORDE, Villancicos 1676–1692)

From this emotionally charged usage, we can surmise that *dizque*, from the beginning of the 17th century, migrates more and more to spoken registers, and it is therefore found more rarely in writing and even in the formal spoken register collected by CDE and CORDE for the 20th century.

4.4 Trends and fluctuations of *dizque* in function of its usage

Looking at the particle *dizque* across the centuries, we can make assumptions about its evolution and usage.

As early as 1944, Kany observed that the form reached its apogee in the first part of the 16th century (168, quoting a 1937 grammar by Keniston) and then slowly declined. However, Kany also maintains that *dizque* did not disappear but "became dialectal, provincial, or rustic" (168)[7] and proceeds to give a number of contemporary examples attesting to the form's spread in many Latin American dialects. Its marginalization in Iberian Spanish was only to be expected given the negative comments found in Golden Age grammars. In Figure 3, the frequency of the *dizque* form is laid out in function of its pragmatic usage:

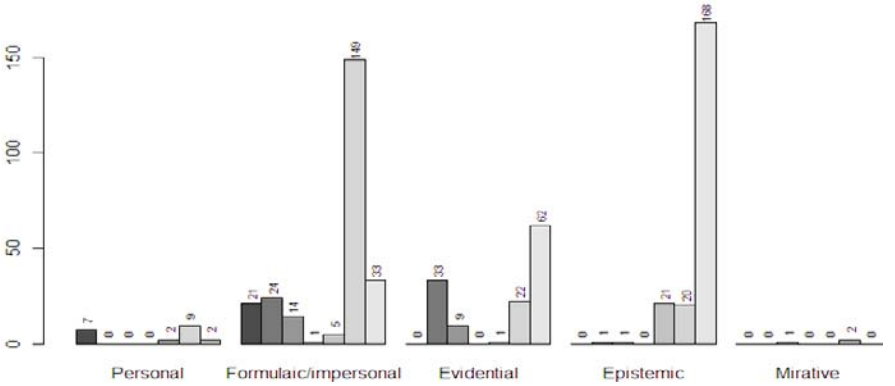


Figure 3. Evolution of *dizque* in function of usage in the CORDE database (darkest column: 1400s, lightest columns are 1900s and CREA).

As could be established from the examples in section 4.1 on the origins of the form, *dizque* starts off as a formulaic collocation 'says + that' (often with the deverbal element separated from its former complementizer). It was originally a personal form 's/he says that' followed by a subordinate clause, but starting from the 1400s (and with Alfonsine prose as a precursor), the form crystallizes into a fixed collocation, more and more frequently written in one word only, with a low agentivity subject (often inanimate, such as 'the story,' 'the document'), or a subject difficult to recover from the context. It is also found translating Latin formulae such as *dicitur* 'it is said' and *vocatur* 'it is called.' These meanings are subsumed under the "formulaic/impersonal" label, and include a usage of *dizque* as *vox populi*, 'it is commonly known, rumor has it.' The step to a reported information marker (i.e., an evidential strategy, simply called "evidential" in the

figure above) is quite short, and one starts seeing such uses in the 1500s.[8] Already in the 1600s the form is in marked decline, and it is all but lost in the 1700s.

Its decline is marked by the emergence of a further extension of the evidential meanings to a marker of surprise (mirative meaning) or disbelief and distancing from the information conveyed (epistemic meaning). These meanings are too emotionally charged to allow an extended use in writing until the 20th century, when typically oral and informal expressions find their way back into the written records. The epistemic *dizque* can be found once again at the end of the 19th century, while the *vox populi* (impersonal), evidential, and epistemic *dizque* all return in great numbers in the 20th century—but mainly in Latin America. Spain must have substituted them with alternative strategies such as the conditional tense or the impersonal *se dice que*, *dicen que* (see note 6 above), which, however, do not fulfill all the roles that *dizque* currently plays in Latin American Spanish (such as the epistemic or the mirative).

In Figure 4 below (the darkest column represents the 1400s and the lightest the 1900s), one can appreciate the same kind of trend in the data gathered from the CDE, specifically the fact that *dizque* acquired new life in the 20th century as an impersonal and epistemic marker. Given that the CDE is about a third of the size of the CORDE, it is not surprising that the actual numbers are somewhat different, especially since the form is not overly common throughout the history of Spanish, and it is perhaps most common in oral, colloquial registers, which are not equally represented compared to other registers in predominantly written corpora such as the CORDE and CDE.

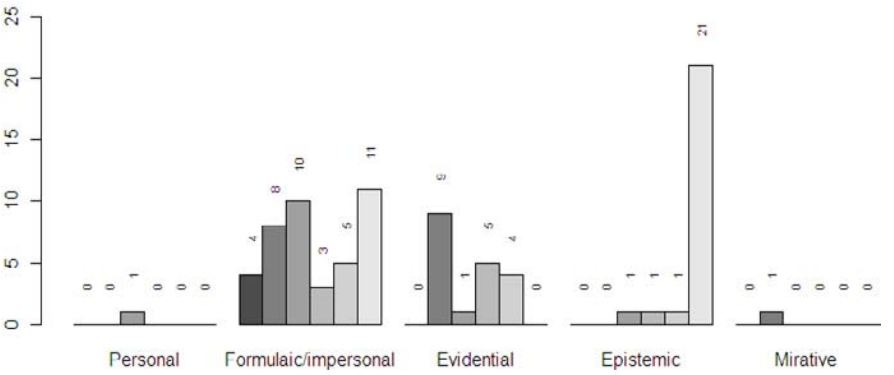


Figure 4. CDE data for *dizque* according to its function through time.

As for Company Company (1994), the evolution of the *dizque* form is easily summed up; this corpus is about 270,000 words and concentrates on Colonial Spanish (from the 1500s to the beginning of the 1800s). It is therefore very small compared to CDE and the RAE corpora, and not surprisingly, there are only few instances of *dizque*. However, because the editor makes a point of gathering documents more likely to contain direct speechlike passages (legal complaints, for instance), one finds three instances of *dizque* at the beginning of the 16th century (two documents from 1529). In all three instances, *dizque* is used

as an evidentiality strategy for facts that the speaker reports without firsthand experience. There are also 15 instances of *diz que*, 13 in the first half of the 16th century and two in the second half; 10 of these occurrences are of the evidential and five of the impersonal (*vox populi*) type.

4.5 The evolution of *dizque* and text types

There is undoubtedly a relation between the spread and subsequent decline of *dizque* on the one hand and the genre in which it was used on the other hand. In its formulaic and impersonal meaning, the usage of *dizque* was born into historical and legal prose (the precursor being the *diz que* form in Alfonsine prose in the 12th century), which is where it abounds in the first two centuries considered in Figure 5 below:

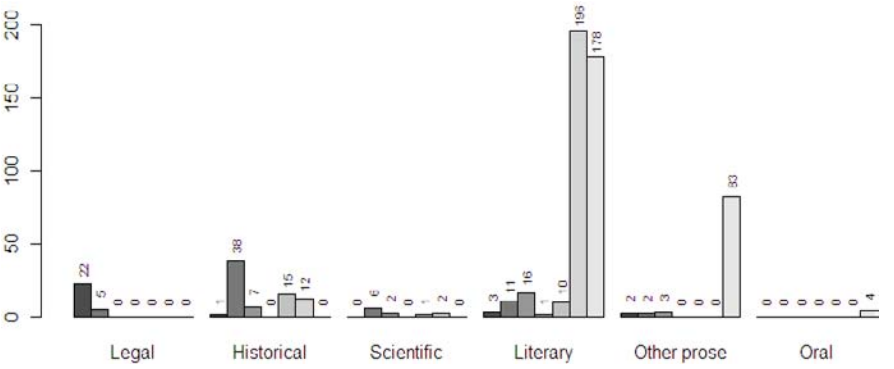


Figure 5. The evolution of *dizque* as a function of text type (register).[9] Darkest column: 1400s, lightest: CREA.

In the 1500s, the *dizque* form also expanded to literary texts, a tendency that prevails throughout the 17th century as the marker of evidentiality begins to acquire negative and may therefore no longer be suitable to the serious prose of a legal or historical text, much less a scientific one. In scientific, legal, and historiographic texts, statements marked as "hearsay" have lost their value and would therefore no longer be found in these registers.

On the other hand, as fiction began to exploit even more the way people actually spoke, such statements became useful in literature as strategies for expressing evidentiality and marking emotively charged speaker evaluations such as disbelief or distance. These developments can explain the *dizque* boom in the 20th century. In fact, we find *dizque* more in narrative than ever before, implying that the form is current in Latin American parlance. This corresponds to anecdotal accounts of *dizque* found in common use in Latin America, but not in Spain. In Spain the form sounds rather foreign (except in Galicia, of course), and consequently, it is not recorded even in writing. This supports Kany's (1944) claim that *dizque* never disappeared completely from Spanish. After having been

stigmatized as rustic speech in Spain, its usage became regional or rural. Latin American corpora might indicate that *dizque* declined throughout the 17th century and was lost in the 18th century, but it is far more likely that it simply disappeared from the written record, possibly because of emotively charged nuances inappropriate for the written register of the time. Based on its persistence in contemporary speech and writing of Latin America, however, one can surmise that *dizque* did survive in the oral register. With the rise of literary channels such as literature and the modern press,[10] in which the epistemic use of *dizque* in writing was appropriate, we find an overall rise in the frequency of *dizque* in written records.

The form is not well represented in the oral component of the CREA corpus, with only four examples. This simply reflects the peculiarities of the corpus and its contents. Both colloquial and formal genres of speech are included in CREA, but the colloquial oral corpora included are dominated by Iberian Spanish, where we would not expect to find *dizque*.[11] Furthermore, the oral part of CREA is small in size compared to the written part of the corpus (10% of those 160 million words), and the Latin American part is made up of political speeches and radio and television texts that are formal and planned, sharing generic features with written prose, where *dizque* is not expected to occur.[12] Even so, we can see that there is an increase in the use of the form in the 20th century in the written part of CREA.

A similar observation can be made for the CDE, which is also a predominantly written corpus, and although its oral part is even smaller than the oral component of the CREA (5.7 million words in CDE vs. 16 million in CREA), we find 12 cases in the CDE oral, as opposed to the four in CREA (see Figure 6 below).

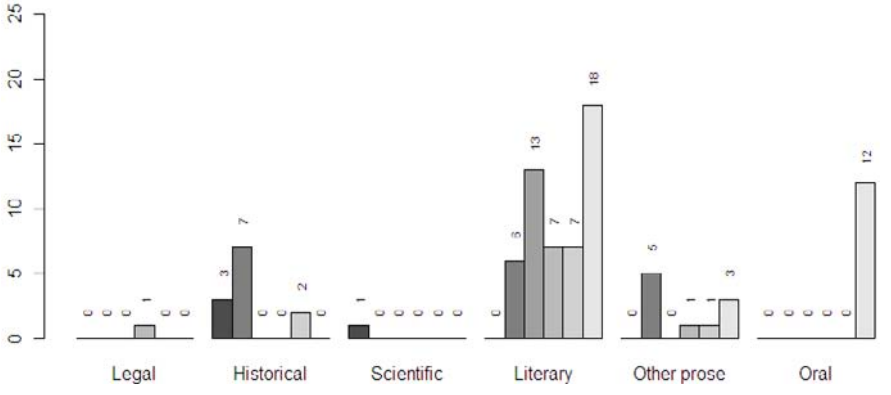


Figure 6. CDE and text types (columns represent 1400s–1900s).

As for Company Company's (1994) small Colonial corpus, all the *dizque* and *diz que* forms come from five documents, four of which are *cartas de relación*, or reports about the state of affairs in the New World to various bodies in Spain, and one of which is a claim to establish residence in the New World.

These documents are examples of historical prose, and align with the findings in other corpora that *dizque* was used in the 16th century as an evidential strategy or as an impersonal quotative reference (*vox populi*).

5. Conclusions

The form *diz que* evolved from the collocation of a 'say' verb plus a complementizer. Resulting from high-frequency occurrence in low-agentivity and quasi-impersonal contexts, *dizque* acquired the function of an evidentiality strategy and grammaticalized into an adverbial particle. From its very inception as an impersonal form, it is clear that *dizque*'s fortunes were connected to specific literary genres such as juridical and historical prose and legal documents.

The usage of the *dizque* increased in the 15th and 16th centuries, and goes hand in hand, again, with the popularity of certain genres such as historical prose, geographical description, and the *cartas de relación.* The use of *dizque* in the written register started to decline in the 17th century, as documented in the CORDE, CDE, and DLNE. This is due in part to the decline of those literary genres in which *dizque* had been used, as well as to drastic changes in historiographical methods, leading to the exclusion of hearsay and unsubstantiated secondhand information as inappropriate sources for the historical record.

The decline of *dizque* in written usage was also caused by factors internal to its semantics. At the beginning of the 17th century, its use in specific contexts changed the form from an evidential strategy to an epistemic marker of disbelief of reported information. And at the same time, one finds instances of *dizque* used as a mirative marker. These emotively charged uses of *dizque* transformed its pragmatic function. As an epistemic marker of speaker evaluation rather than an evidential marker, *dizque* migrated to the oral register as well as prose portraying spoken forms such as fiction and the modern press. This is where we find *dizque* in corpora of literary genres once again, alive and kicking so to speak, in the 20th century.

The impressionistic view of speakers of Latin American Spanish LAS that the form is very common in oral usage is very likely true, but hard to corroborate with data from corpora, whose LAS component usually comprises mainly formal spoken Spanish data (*lengua culta*). Its frequency in oral LAS is most likely reflected in modern and contemporary fiction closer to the spoken register. My suggestion is that *dizque* changed registers starting from the 16th century and became a predominantly spoken or colloquial strategy. This would explain why the form remains rather scarce in the later CDE and CORDE documents, although its spike in modern Spanish is clearly attested in fiction, a register well represented both in CORDE and CDE.

Both the main online corpora CDE and CORDE/CREA correctly reflect the usage and evolution of the form, as can be expected given the abundance of collected. The fact that *dizque* seems more common in modern Spanish according

to CDE (1.4 occurrences per million) than in CREA (0.00000166 per million) may depend on the choice of documents included in the respective corpora (more or less formal, from Latin America or from Spain), and may superficially skew the data referring to this specific structure. However, there is a remarkable correspondence between the data in CDE and CORDE/CREA as far as general tendencies go, since both corpora reflect the same trends in the semantic evolution for *dizque* and its contexts. Significantly, even a much smaller corpus of documents from the Mexican Colonial Period (Company Company, 1994) reflects a similar trajectory of evolution. The agreement among these very different corpora indicates that the semantic changes proposed here for *dizque* and its distribution in different text types indeed reliably reconstruct the historical development of this form.

It is clear that the ease with which we can corroborate the historical evolution of a linguistic form would be transformed in painstaking library or archival work without the help of large online corpora such as CDE and CORDE/CREA. Given the differences between the corpora established in this paper, archival work would be much more flawed without the access to the enormous amount of data provided by the online corpora, which are indeed an invaluable tool for diachronic as well as diatopic and diastratic linguistic research.

6. Notes

* I am greatly indebted to Stefan Th. Gries and two anonymous reviewers for many helpful comments and suggestions that have greatly improved this paper. All remaining errors and infelicities are of course my own.

1. All translations of Spanish examples are mine unless otherwise specified.
2. The bulk of the work on the corpora was carried out in January and February of 2008; the composition of all three main online corpora (CDE, CORDE, and CREA) has changed since then. I have tried to minimize numeric discrepancies from later corpus usage and revisions to this article, but some undoubtedly remain.
3. <http://www.rae.es/rae/gestores/gespub000019.nsf/voTodosporId/F651954693D3AC5EC125716400426004?OpenDocument> (accessed on August 1, 2008).
4. Omar Miranda and Elizabeth Ortega (p.c.), native speakers of Mexican and Colombian Spanish, respectively.
5. A quick perusal of these forms by region in RAE shows that the CREA data do not show consistent differences. *Se dice que* is used 833 times in LAS and 909 times in Spain, but *dicen que* is used 2,865 times in LAS and 2,406 times in Spain. Twentieth-century data in CORDE, on the other hand, seem to corroborate that these forms are used more often in Spain than in Latin America: *se dice que* is used only 112 times in LAS and 705 in Spain, and *dicen que* is used 612 times in LAS and 1,103 times in Spain. A more refined search for equivalent evidentiality strategies

remains to be carried out, since some of them could consist in the use of certain tenses, for instance the conditional (such as in *los ladrones habrían salido del banco sin que nadie los viera*, 'the thieves supposedly left the bank without being seen'), and not in lexical strategies such as the one analyzed here (accessed on December 1, 2008).

6. Which could simply refer to the document (i.e., 'the document says that...').
7. López-Izquierdo (2005) also agrees that the adverbial form is still in use in Iberian Spanish, if only rarely.
8. It is admittedly difficult to distinguish between impersonal and evidential usages at times, and to a certain degree it is left to the researcher's discretion. I have tried to limit labeling examples of *dizque* as evidentiality strategies to instances when the act of reporting and transmitting information was clear from the verbal forms or from the extended context, avoiding *vox populi*-type of meanings that would be collected under formulaic/impersonal.
9. Numbers between the two figures may not always coincide. At times (for instance in the 1500s), CORDE counts 62 cases but only 58 can be recovered and analyzed, and in the 1700s, CORDE counts two examples, but it can only show one. Only analyzable examples were counted for Figure 3, but in Figure 5, the statistics were done automatically by CORDE, and all the examples it counted are to be found there, even those I could not analyse because they could not be displayed.
10. Classified as "other prose" in the CREA columns.
11. The makeup of the CREA corpus can be found at <http://www.rae.es/rae/gestores/gespub000019.nsf/voTodosporId/F651954693D3AC5EC125716400426004?OpenDocument>.
12. Further discussion of the ties between *dizque* and genre can be found in López-Izquierdo and Miglio, 2008.

References

Aikhenvald, A. Y. (2004), *Evidentiality.* Oxford: Oxford University Press.

Aikhenvald, A. Y., and R. M. W. Dixon (2003), *Studies in evidentiality.* Amsterdam/Philadelphia: John Benjamins.

Company Company, C. (1994), *Documentos lingüísticos de la Nueva España.* Altiplano-Central, Mexico: U.N.A.M.

Company Company, C. (2006), "Subjectification of verbs into discourse markers: semantic-pragmatic change only?" *Belgian journal of linguistics,* 20(1): 97–121.

Covarrubias Orozco, S. de (1995 [1611]), *Tesoro de la lengua española o castellana.* Madrid: Castalia.

Davies, M. (2002), *Corpus del español*. Online at: <http://www.corpusdelespanol.org>.

Lapesa, R. (1981), *Historia de la lengua española.* Madrid: Gredos.

López-Izquierdo, M. (2005), "L'émergence de dizque comme stratégie médiative en espagnol médiéval," *Cahiers de linguistique et de civilisation hispaniques médiévales*, 29.

López-Izquierdo, M., and V. Miglio (2008), "A multi-factorial approach to grammaticalization and lexicalization: the case of Spanish *dizque*," paper presented at the 4th conference new reflections on grammaticalization, Leuven, Belgium.

Magaña, E. (2005), "El paso de 'dice que a dizque, de la referencia a la evidencialidad," *Constribuciones desde Coatepec,* 8: 59–70.

Menéndez Pidal, R. (1987), *Manual de gramática histórica Española*. Madrid: Espasa Calpe.

Miglio, V. (in press), *Tengo por cierto*, *hera fama que*, *diziendo que*, *dizque: referencias a la realidad en el español de la Colonia.* Proceedings of the 9th UC-Mexicanistas colloquium on Mexican culture and literature in honor of Prof. Timothy M. McGovern, UC Santa Barbara.

Nebrija, A. de (1492), *Gramática de la lengua castellana*. Online at: <http://www.antoniodenebrija.org>.

Real Academia Española (1726–39), *Diccionario de la lengua castellana*. Madrid, F. del Hierro. [Ed. facsimile, 1963]: Madrid, Editorial Gredos, 3.

Travis, C. E. (2006), "*Dizque*: a Colombian evidentiality strategy," *Linguistics,* 44(6): 1269–1297.

Toward a comparison of unsupervised diachronic morphological profiles

Alfonso Medina Urrea

Instituto de Ingeniería, Universidad Nacional Autónoma de México

Abstract

Very diverse, unsupervised methods exist for segmenting graphical words. Some of these can be applied to compile sets of affixes and their sequences (i.e., morphological profiles). These sets seem to intimately characterize the corpora of a variety of languages. A general measurement of variation at a morphological level can be obtained by comparing different profiles of one given language (e.g., from different diachronic states), and obtaining distances that can be used for corroborating or discovering the dialectal structure of concatenative languages. In this paper, we present quantitative data from corpora of three centuries of the Spanish language used in what is today Mexico (16^{th}, 18^{th}, and 20^{th} centuries) and small samples of Peninsular Spanish. Euclidean distances between these morphological profiles are calculated and discussed. Thus, some intuitions concerning the morphological level are preliminarily substantiated, like the supposed emergence of the Mexican Spanish dialectal system sometime before the 18^{th} century.

1. Introduction

It is no secret that through using corpora, many things can be learned about languages in terms of systems. From the fundamental work of linguists like Zellig Harris to the practical applications of today's text mining, phenomena at all linguistic levels are there to be discovered or extracted. Morphology represents a linguistic level of interest to both linguists and text miners. Hence, many unsupervised procedures were developed and tested from the fifties onwards in order to discover morphemes. In this paper, I briefly describe a simple method for extracting affixes and their sequences from corpora. It has been previously applied to discover certain affix sets of concatenative languages (Medina and Buenrostro, 2003; Medina and Hlaváčová, 2005; Medina and Alvarado, 2006; Medina, 2007) like the derivational suffixes of Ralamuli or Tarahumara (precision/recall 0.80/0.71); Czech prefixes (precision/recall 1.0/0.80); and the prefixed and suffixed verbal inflection of Chuj, a Mayan language (prefixes 0.75/1.0; and suffixes 1.0/0.75). As may be appreciated, this method can be applied to compile unsupervised catalogs of these items, at least in the case of concatenative languages such as Spanish. These catalogs include statistical values grading each item's likelihood of being used as an affix in that language. In fact, it appears that they can be considered actual morphological profiles of the corpora from which they are extracted (as opposed to plain sets of strings unrepresentative of morphological items). If this is the case, morphological extractions from

different corpora representing different dialects (diachronic or synchronic) deserve to be compared. For this purpose, simple distance measurements are presented and discussed. In short, the idea behind this experiment is that certain morphemes that characterize a given corpus can be extracted automatically and later compared to the morphemes extracted from some other corpus representing another dialect of the same language, thus revealing the distance between them. These distances provide a simple method for corroborating or even discovering the dialectal structure of concatenative languages at the morphological level.

2. Background

Embleton (1986) refers to the great amount of quantitative work that has been undertaken in order to assess the degree of genetic relatedness between languages with the aim of comparing them diachronically. Swadesh's (1955) glottochronology represents by far the most prominent approach—it requires the compilation of lexical bases of at least 100 items (body parts, heavenly phenomena, small numerals, personal pronouns, basic action verbs, etc.) in order to compare cognates from different languages and calculate their separation from the ancestral language in terms of millennia. It is important to recall that *cognates* consist of words with a common origin, related by descent from the same ancestral language. They may exist between languages—e.g., English *starve* and German *sterben*—or within languages—e.g., English *shirt* and *skirt*; and Spanish *delicado* and *delgado*. Also, in relation to this paper's objectives, morphemes within words may constitute cognates—e.g., the verbal clitics of Romance languages. I will argue that these latter types of cognates may constitute a crucial basis for the unsupervised comparison of languages.

When dealing with genetically related languages, several factors exist indicating that a comparison of morphological rather than lexical items may be relevant. Firstly, lexical items, specifically the root morphemes within them, convey most information in the situations in which they are used. Secondly, some morphological items, specifically modifiers, clitics, and affixes (which are inflectional and to some extent derivational) express the grammatical information that structures discourse. Contrastingly, a list of basic, lexical items—such as that presupposed in glottochronology—represents what can be referred to as *culture* rather than a grammatical system of communication that gives linguistic structure to human thought and culture.

It is true that items in a compiled lexical base may also be parts of the grammatical system, inasmuch as they function as pronouns, classifiers, or numerals (functions typical of items within grammatical systems). If they function as such, they belong to the set of grammatical modifiers or items of clitical or affixal nature (i.e., frequent items that have been stripped of much of their phonological material and their full semantic or referential meaning). Furthermore, in dealing with genetically related languages, these are very likely to share morphological cognates of this sort. The more distant these languages

are, genetically or even geographically, the less likely they are to share such structural items.

Needless to say, in order to compare less-related languages, the manual compilation of customary lexical bases will be more compelling. Nevertheless, such comparison will still reflect less the level of an intimate communication structure (which unrelated languages do not share) than that of culture and thought. This is not to say that a list of items represents one culture or all cultures; nevertheless, the fact that several languages may share referents does not make their signifiers culture-independent. And when these signifiers refer to supposedly universal concepts, they rather characterize the intersection of sets of culture-relevant referents.

Another interesting point refers to the affixal nature, or *affixality*, of certain morphological items. Graphical word fragments may be conceived as more or less affixal depending on the extent to which they may be joined to other items in order to form other graphical words, as attested in a corpus. Besides this, the affixal quality of word fragments may be estimated from the corpus. Using these estimations, affix candidates may be ordered, going from the most to the least affixal, in tables called *morphological profiles.* This affixality may be regarded as an example of what is termed *glutinosity* (Medina Urrea, 2003), defined as the glue that sticks morphological items together and that varies across time and space. Glutinosity may be regarded as Sapir's (1921: 47) energy of sequences:

> Words and elements, then, once they are listed in a certain order, tend not only to establish some kind of relation among themselves but are attracted to each other in greater or in less degree. It is presumably this very greater or less that ultimately leads to those firmly solidified groups of elements (radical element or elements plus one or more grammatical elements) that we have studied as complex words […]. Speech is thus constantly tightening and loosening its sequences. In its highly integrated forms (Latin, Eskimo) the "energy" of sequence is largely locked up in complex word formations and becomes transformed into a kind of potential energy that may not be released for millennia. In its more analytic forms (Chinese, English), this energy is mobile, ready to hand for such service as we demand of it.

Thus, affixality can be regarded as the glutinosity between affixes and bases. In this experiment, affixality will be defined as the force that glues morphemes in terms of bits (entropy) carried by some kind of structure (economic squares); these concepts will be described below.

3. Morphological profiles

There are many techniques for morphological segmentation. Some prominent and recent ones include bigram statistics (e.g., Kageura, 1999), minimal distance methods (Goldsmith, 2000), and Bayesian statistics (Creuz and Lagus, 2005). The method I will discuss below may be regarded as an elaboration of Zellig Harris's approach (1955)—where uncertainty is measured by counting phoneme varieties—and Josse de Kock's economy principle (1978)—as the structural nature of lexical items. There is a lack of space here for disputing the advantages of one method over the other, so in the following paragraphs, I will briefly describe this method's three criteria for determining affixes (counts of squares, entropy, and economy values).

3.1.1 Counts of squares

An important indication of the validity of segmentation is the notion of *square*. Joseph Greenberg (1957: 20) characterized it succinctly as the set of strings that

> exists when there are four expressions in a language that take the form AC, BC, AD, BD. An example is English eating:walking::eats:walks, where A is eat-, B is walk-, C is -ing, and D is -s. One of the four members may be zero, as in king:kingdom::duke:dukedom, where C is zero.

Thus, a square will be a set of four word fragments, two beginning ones (a_1 and a_2) and two ending ones (b_1 and b_2), such that when combining either of the former with either of the latter, a word-type results (a_1b_1, a_1b_2, a_2b_1, or a_2b_2) that is an element of the set of word types of the corpus. Thus, each possible segmentation of each word-type i may be examined mechanically, in order to determine the number of squares it yields, by searching for matches in the set. In other words, a computer program goes through the corpus, splits up every word into two parts in all possible ways, and, for each hind part of a word, determines how many squares can be filled with all other possible front and hind parts of words. Let us call this number $c_{i,j}$ (i.e., the number of squares found in segment j of the word type i).

Table 1. Number of squares for each segmentation of word type *aumente* (20th century Mexican Spanish).

A	**U**	**M**	**E**	**N**	**T**	**E**
	0	1,021	20	0	0	**8,348**

For example, Table 1 displays the number of squares found for each cut of the word type *aumente.* The highest number corresponds correctly to the border between the root *aument-* and the subjunctive suffix *-e*.

3.1.2 Information content

Information content of a set of graphical words is a widely recognized concept and is typically measured using Shannon's entropy (1949), embodied in the well-known formula

$$H(p_1, p_2, ..., p_n) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

where p_i stands for the relative frequency of word fragment *i*. High entropy measurements of word fragments have been repeatedly reported as representing successful indicators of borders between bases and affixes. For this experiment, the task was to measure information content in the set of all word fragments, occurring with each possible suffix candidate. Thus, given a word-type, with entropy values calculated for each suffix candidate, the morphological border will most probably correspond to the highest entropy value. Let $h_{i,j}$ be the entropy measurement of word fragment *j* of word type *i*.

Table 2. Entropy measurements (bits) for each segmentation of *aparec|ser* (20th century Mexican Spanish data).

	A	**P**	**A**	**R**	**E**	**S**	**E**	**R**
left-right		**2.792**	1.818	1.63	1.298	1.27	0.9497	1.303
right-left		1.277	0.8018	1.619	2.125	1.56	**2.516**	1.193

Table 2 displays the entropy expected right after the beginning of the word (from left to right) and information content of everything that precedes the possible word endings (from right to left). As it can be observed, the first line provides data to select prefixes, while the second line shows criteria for suffixes. In this case, the best prefix is *a-* and the best suffix is *-er*.

3.1.3 Economy principle

Concerning the syntagmatic dimension, affixes combine with bases to produce a (virtually infinite) number of lexical signs. It makes sense to expect more economy where more combinatory possibilities exist. Concerning the paradigmatic dimension, affixes occur in complementary distribution with other affixes. The economic nature of these dimensions can be measured by comparing the following sets of fragments of word types. Given a suffix candidate, there are at least two sets of interest:

- *Companions*—word beginnings that are followed by the suffix candidate in syntagmatic relation; we can call this set *A*.
- *Alternants*—word endings that occur in complementary distribution with the suffix candidate (i.e., they alternate with it in that position); we can call this set *B*.

The ratio of these sets' sizes is a simplified method of capturing the economy associated with a suffix candidate. More formally, let $A_{i,j}$ be the set of word beginnings that is followed, according to a corpus, by the right-hand word segment $b_{i,j}$ (which is the string of letters beginning at the *j*th column of the word type *i*). Let $A^p_{i,j}$ be the subset of $A_{i,j}$ consisting of the word beginnings that behave like prefixes of the language in question. Let $B^s_{i,j}$ be the set of word endings that behave, also according to the corpus, as suffixes of the language and occur in complementary distribution with $b_{i,j}$. One way to estimate the economy of a segmentation between a set of word endings and a set of word beginnings, in such a way that the word endings are suffixes is

$$k^s_{i,j} = \frac{\left|A_{i,j}\right| - \left|A^p_{i,j}\right|}{\left|B^s_{i,j}\right|}$$

The numerator of this formula can be described as the set of left-hand companions of the right-hand word segment $b_{i,j}$, and the denominator is the set of right-hand segments or alternants of (in paradigmatic relation to) $b_{i,j}$. In this way, when a word fragment is given, a very large number of companions and a relatively small number of alternants yield a high economy value. Meanwhile, a small number of companions and an ample number of alternants indicate a low economy measurement. In the latter case, the word fragment in question is not very likely to precisely represent a morpheme or a set of these.

Table 3. Economy indexes for each segmentation of word type *comente* (20th century Mexican Spanish data).

C	O	M	E	N	T	E
	0	0	0	0.476	0.940	**0.945**

For example, Table 3 displays the economy measurements obtained at each possible cut of the word *comente.* The highest value corresponds correctly to the border between the root *coment-* and the inflection *-e*.

3.2 Squares, entropy, and economy combined

The measurements described above complement each other concerning the estimation of the affixality of word fragments. The values obtained for a given word fragment can be averaged or multiplied. For this experiment, the three values were calculated for each possible segmentation, for each word type, and for each corpus, and these were then normalized and averaged. Thus, affixality was estimated by considering the arithmetic average of the relative values of counts of squares, entropy, and economy:

$$AF_i^s = \left(\frac{c_i}{\max c} + \frac{h_i^s}{\max h} + \frac{k_i^s}{\max k} \right) * \frac{1}{3}$$

where c_i stands for the normalized number of squares attested for affix candidate i; $h^s{}_i$ stands for the information content of all word fragments occurring before the suffix candidate i; and $k^s{}_i$ represents the economy measurement associated with candidate i. Since both entropy and economy are directional values (i.e., two values exist at each segmentation, one representing a possible prefix and one a possible suffix), the last two values are tagged with a superscript indicating suffixality. It is important to note that the highest values of AF^s, those expected to occur at the borders between bases and suffixes, represent very good criteria for choosing word fragments to be included in the suffix catalogues. Henceforth, the procedure for building a catalog basically consists of taking each word type, determining the best suffix, and inserting it in the catalog, along with all the measurements obtained.

Finally, regarding the three measurements by means of which AF^s is estimated, peaks of entropy select sequences of affixes attached to roots, including old and unproductive derivations; the economy index selects the outer affixes of words, mainly inflections; and the squares index identifies chained sequences of morphemes, not necessarily affixal (e.g., bases belonging to compositions plus accompanying derivations and inflections). In previous experiments with 20th century Spanish, it was observed (Medina, 2003) that each method obtained different precision rates: 87.22% (squares), 86.39% (entropy), and 79.17% (economy); whereas the conflation of these three measurements obtained a precision rate of 95% (Medina, 2007). In one sense, these three measurements constitute a voting committee selecting the best segmentation. In another sense, they are magnitudes of independent dimensions expressing a new, derived dimension that we tentatively call affixality or, in a wider sense, glutinosity. In other words, the idea is to derive from these magnitudes a new magnitude, in an analogous manner to how a velocity magnitude is derived from distance and time.

4. Morphological profiles for three stages of Spanish in septentrional America

As mentioned above, the language states selected for this experiment correspond to the 16th, 18th, and 20th centuries of Spanish written in Mexico. For the first two stages, we used documents from the *Corpus Histórico del Español en México* (CHEM), currently being compiled at our Institute. For the 20th century, we used data previously extracted from the *Corpus del Español Mexicano Contemporáneo* (CEMC) compiled at El Colegio de México, using basic criteria designed to fulfil representativity (Lara, 1974)—this consists of paragraphs randomly picked from almost a thousand documents written between 1921 and 1974, and grouped into a

number of categories of cultivated speech (literature, journalism, science, etc.), popular speech, and nonstandard speech (slang, specialized registers, anthropological documents, etc.). Thus, it is important to observe that these samples are very uneven concerning their representativity.

Once the method described above was applied, suffixes and suffixal groups were extracted from each corpus and stored in catalogs for later comparison. The data appear interesting; however, some caution is required regarding the nature of the samples used.

For example, one important aspect when comparing texts from different dialects concerns the desirability of phonological transcription. However, this task presents several difficulties. Firstly, Spanish phonology underwent important changes around the 16th century, so specific transcription rules for that era are required. Some rules have been proposed for the automatic transcription of 16th century documents of the CHEM (Reyes Careaga, 2008), but they are still very tentative, and consensus seems far from being reached. Also, even if this were solved, the great orthographic irregularities of old documents make automatic phonological transcriptions extremely difficult. Therefore, this experiment is basically a written language exercise, even though we are dealing with a language with a high degree of spelling/pronunciation correlation.

Fortunately, when the textual samples and the extracted suffix sets are examined, it is possible to observe that the great orthographic variability in old documents occurs mostly in the roots and bases of graphical words. Also, since the documents were edited following the philologists' customary practice of reconstructing abbreviated graphical words, when affixes and affix chains were reconstructed, these were standardized, limiting the orthographical variability of affixes in reconstructed abbreviations.

The results are displayed in Tables 4, 5, and 6, which show the first 10 morphological items for each century, in order from the most affixal (exhibiting more glutinosity) to less affixal (exhibiting less); smaller ranks imply greater affixality.

Table 4. Spanish suffixal groups from the 16th century (CHEM); first 10 of 760 entries of catalog for this century.

#	Suffix	Frequency	Squares	Economy	Entropy	Affixality
1	a	919	0.6410	0.9441	0.9505	0.8452
2	s	1,288	1.0000	0.9968	0.4833	0.8268
3	o	902	0.6676	0.9355	0.8164	0.8065
4	ó	306	0.6420	0.8362	0.8720	0.7834
5	as	403	0.3457	0.9325	0.9607	0.7463
6	ar	259	0.3003	0.9447	0.9623	0.7358
7	os	507	0.3721	0.9253	0.8883	0.7286
8	ado	250	0.2681	0.9434	0.9631	0.7249
9	e	534	0.4770	0.8904	0.7931	0.7202
10	aba	137	0.1979	0.9093	0.9415	0.6829

Thus, Table 4 presents the first most suffixal items out of a catalog of 760 entries, extracted from a sample of 151,966 tokens and 17,608 types. Table 5 presents those from a catalog of 527 entries extracted from a sample of 165,159 graphical word tokens and 15,916 word types.

Table 5. Spanish suffixal groups from New Spain's 18th century (CHEM); first 10 of 527 entries of catalog for this century.

#	Suffix	Frequency	Squares	Economy	Entropy	Affixality
1	a	1,462	0.6558	0.9564	0.9371	0.8497
2	o	1,418	0.7031	0.9574	0.8153	0.8253
3	s	1,761	1.0000	1.0000	0.4558	0.8187
4	as	593	0.3461	0.9503	0.9331	0.7432
5	ó	287	0.4709	0.8447	0.9094	0.7417
6	os	657	0.3799	0.9460	0.8863	0.7374
7	ar	353	0.2279	0.9533	0.9658	0.7157
8	ado	341	0.2184	0.9566	0.9555	0.7102
9	e	629	0.3629	0.8919	0.7683	0.6744
10	an	285	0.1799	0.9059	0.9235	0.6698

Evidently, the samples from which these two affix sets were extracted are rather small. Contrastingly, Table 6 presents the 10 most affixal items out of 749 affixes and groups of affixes extracted from the 20th century corpus of about two million word tokens and almost 70,000 types of words.

Table 6. Spanish suffixal groups from Mexico's 20th century (CEMC); first 10 of 749 entries of catalog for this century.

#	Suffix	Frequency	Squares	Economy	Entropy	Affixality
1	ó	1,428	0.7371	0.9192	0.8720	0.8428
2	o	6,314	0.6860	0.9788	0.8017	0.8222
3	s	12,013	1.0000	0.9968	0.4609	0.8192
4	a	7,687	0.5753	0.9818	0.8888	0.8153
5	os	4,554	0.4775	0.9754	0.8235	0.7588
6	as	4,324	0.4216	0.9779	0.8645	0.7547
7	en	945	0.4107	0.8991	0.9060	0.7386
8	ar	1,633	0.2178	0.9621	0.9149	0.6982
9	ado	1,429	0.2061	0.9619	0.9070	0.6917
10	ando	976	0.1836	0.9544	0.9162	0.6847

Regarding the items extracted, it is worth observing that some constitute groups of suffixes. Take for instance the item *-os*, which encapsulates two inflectional morphemes: noun gender (*-o*) and number (*-s*) markers. These structures do not pose a problem for this experiment, since sequences of this kind are common in many languages. In fact, affixal structures like this should be

expected to appear in tables like these if they are typical of the language and happen to be more affixal than other isolated items. One might think of those Sapirean sequences locked in complex affixal groups, which emerge as unified affixal items during different stages of the same language. These sequences certainly deserve special attention if one is to discover the affitactics of the language.

It is interesting to note that two suffixes of similar affixality may be, in fact, very different from each other. For example, suffix *-s* is high on squares and economy but relatively low on entropy, while suffix *-ar* scores high on economy and entropy but very low on squares (both among the top 10 of Tables 4, 5, and 6). As mentioned above, entropy extracts well the sequences of affixes attached to roots, while the economy index extracts the outer affixes of words, mainly inflection. Thus on the one hand, an affix with high squares and economy scores combined with low entropy would be characteristic of the outermost inflections, typically attached to other inflections, which is the case of suffix *-s*. In one of its meanings, it attaches to inflected nouns, which are farthest from the root, to form plurals (e.g., sólid-o-*s*, erudit-a-*s*, prend-id-it-o-*s*, canast-ita-*s*). On the other hand, an affix with high entropy and economy values but low squares counts would be more likely an inflection that attaches to bound roots, like suffix *-ar* which attaches directly to roots of verbs, and are not free morphemes (e.g., compr-*ar*, naveg-*ar*, alegr-*ar*). This analysis is not precluded when the scores are conflated to estimate affixality for unsupervised affix extraction. Failing to see the utility of estimating affixality seems somehow analogous to failing to see the use of calculating velocity when trying to determine the fastest animals of a particular zoo, of a bird flying 30 kilometers in 10 hours, and of a zebra rambling 6 kilometers in 2 hours: similar velocity, different moving agents, different contexts.

Finally, although a certain similarity between these three profiles is evident, there are also some differences: Tables 4 and 5 share 9 out of 10 items with slight variations in their order. Table 6 shares eight items with the first two profiles and also exhibits some variation in the order. It is nevertheless very hard to assess the distance between them, especially if we are to consider their affixality values.

4.1.1 Distances between morphological profiles

A variety of measurements exists that are applicable to the comparison of these profiles (i.e., other distance coefficients such as Minkowski, Manhattan, or Canberra distances) (Oakes, 1998). Also, other coefficients can be used that measure degree of correlation rather than distance, such as Pearson's r or Spearman's ρ (see also Oakes, 1998). However, for this experiment, the simple Euclidean distance was preferred for this presentation because of its simplicity and straightforwardness. Thus, distances between each pair of profiles were measured using the formula

$$d_{jk} = \sqrt{\sum_{i=1}^{n} (X_{ij} - X_{ik})^2}$$

where X_{ij} stands for the affixality value of the morphological item i in profile j, and X_{ik} stands for the affixality value of the same item in morphological profile k; n is the number of items they share. The averaged values are obtained by applying the following formula:

$$\delta_{jk} = \sqrt{\frac{\sum_{i=1}^{n} (X_{ij} - X_{ik})^2}{n}}$$

Consequently, the averaged Euclidean distances calculated for all items of the morphological profiles partially presented above appear in Table 7. Each cell shows distances using all the shared items and not just those shown in Tables 4, 5, and 6. The diagonal represents the distance of each century from itself.

Table 7. Euclidean distances between some Spanish diachronic dialects used in septentrional America.

	16th	**18th New Spain**	**20th Mexico**
16th	0.0000	0.0781	0.0913
18th New Spain		0.0000	0.0715
20th Mexico			0.0000

It is interesting that these values are very small. However, perhaps this is not surprising, since we are dealing with relatively close diachronic stages of a language known for its conservative nature. The puzzling aspect concerning these small values is that they were obtained from a set of such uneven samples (two very small and a comparatively enormous one). One would expect the questionable representativity of the smaller ones to be reflected here. Does this mean that the morphology of languages such as Spanish can be captured in such small corpora? Or are we dealing here with a random event where three sets of numbers simply happen to be very similar?

At least it appears that the sets of affixes and their sequences are more intimately related to the languages in this experiment than any set of basic lexical items relied upon in glottochronology.

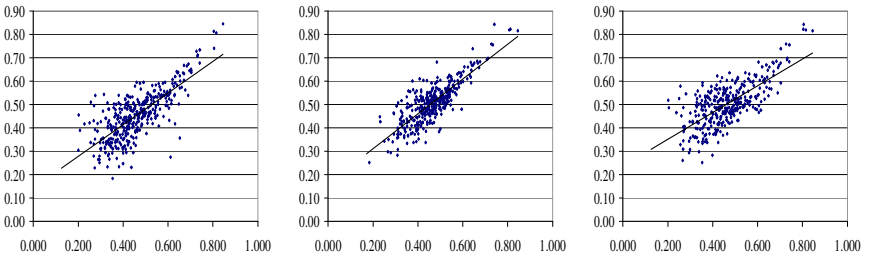
It is also interesting that, as might be expected, the greatest distance occurs between the 16th and the 20th centuries. Also, the 18th century appears closer to the 20th century, corroborating the idea that Mexican Spanish as a distinct dialectal system may have already existed by the 18th century.

Figure 1 displays summary graphs for each pair of centuries. Each item shared by each pair of morphological profiles appears plotted in the corresponding graph. Graph B, corresponding to the 18th and 20th centuries, looks more compact, exhibiting the shortest distance. Graph C, corresponding to the 16th and 20th centuries, looks more disperse (greatest distance). The dispersion in

A, corresponding to the 16th and 18th centuries, is visually greater than that in B, pointing again to the possible emergence of a distinct dialectal system sometime before the 18th century. However, all these observations and statements are questionable given the fact that the distances seem simply too small. These numbers might be put in context by adding other Spanish dialects to our samples.

4.1.2 Distances to Peninsular Spanish profiles

Although the Spanish language includes a number of prominent and prestigious national dialects, for historical reasons, Peninsular Spanish or Castilian was picked for this experiment and also because of the greater availability of textual samples from which to extract morphological profiles.



A (0.0781) B (0.0715) C (0.0913)

Figure 1. Dispersion graphs. Data of 16th and 18th centuries are compared in A (374 shared items); data of 18th and 20th centuries are compared in B (362 items); and data of 16th and 20th centuries are compared in C (335 items). Euclidean distances appear within parentheses after labels A, B and C.

Thus, small textual samples were improvised for the 18th and 20th centuries by searching the *Corpus Diacrónico del Español* (CORDE) and the *Corpus de Referencia del Español Actual* (CREA) maintained by the Real Academia Española. In order to avoid bias, searches were carried out for content words (e.g., *elefante*) and *españolismos* (i.e., content words used mostly in Spain with equivalents in other dialects, e.g., *bañador*), and the resulting concordances were put together as textual samples. The searched documents were produced in Spain. Tables 8 and 9 contain Castilian suffixal groups for the 18th and 20th centuries, respectively.

It is worth observing that Penisular Spanish differs from all American dialectal systems in its phonology. Given that this experiment is based on written texts rather than phonological transcriptions of them, one should expect the distances between the profiles of these experiments to be in some way more conservative (i.e., smaller) than in reality. For example, graphical suffixes containing *z* (like those used in patronymics such as Gonzál-*ez*, Rodríg-*uez*, etc., or those found in other derivations like vej-*ez*, bell-*eza*, etc.) would have different phonological transcriptions depending on the side of the Atlantic (/gon.θá.leθ/ *vs.*

/gon.sá.les/, /be.xéθ/ vs. /be.xés/). This information gets lost when comparing plain text samples.

Table 8. Peninsular Spanish suffixal groups from the 18th century; top 10 of 429 entries of catalog for this century.

#	Suffix	Frequency	Squares	Economy	Entropy	Affixality
1	s	1,546	1.0000	1.0000	0.5012	0.8338
2	a	1,063	0.5585	0.9573	0.9628	0.8262
3	o	1,050	0.5792	0.9448	0.8379	0.7873
4	as	477	0.3156	0.9417	0.9360	0.7311
5	os	586	0.3242	0.9278	0.9002	0.7174
6	ar	245	0.2027	0.9449	0.9743	0.7073
7	ado	221	0.1507	0.9393	0.9724	0.6875
8	an	247	0.1676	0.9016	0.9422	0.6705
9	ando	128	0.1147	0.9016	0.9641	0.6601
10	ó	200	0.3371	0.7885	0.8369	0.6542

As mentioned above, Tables 8 and 9 show the most affixal items out of the samples gathered. The catalogue presented in the former has a total of 429 entries, extracted from a collection of 96,877 tokens, corresponding to only 13,882 types. The affixes in Table 9 represent the first 10 out of 551 items, extracted from a sample of 125,969 word tokens (17,509 word types). Again, the samples are very small and the results appear similar in some way.

Table 9. Peninsular Spanish suffixal groups from the 20th century; top 10 of 551 entries of catalog for this century.

#	Suffix	Frequency	Squares	Economy	Entropy	Affixality
1	s	2,071	1.0000	0.9952	0.5179	0.8378
2	a	1,470	0.5112	0.9617	0.9727	0.8152
3	o	1,398	0.5961	0.9501	0.8692	0.8051
4	as	760	0.3452	0.9486	0.9710	0.7549
5	os	670	0.3196	0.9297	0.9221	0.7238
6	ó	303	0.3869	0.8556	0.9064	0.7163
7	ado	337	0.1668	0.9261	0.9786	0.6905
8	ar	359	0.1830	0.9086	0.9676	0.6864
9	aba	217	0.1423	0.9125	0.9587	0.6712
10	ando	167	0.1149	0.8934	0.9913	0.6665

As before, the affixality values in these profiles were used to calculate the averaged Euclidean distances presented in Table 10 below, where the Peninsular dialects for the 18th and 20th centuries are added.

Table 10. Euclidean distances between some Spanish diachronic dialects.

	16th	**18th New Spain**	**20th Mexico**	**18th Peninsula**	**20th Peninsula**
16th	0.0000	0.0781	*0.0913*	*0.0833*	*0.0877*
18th New Spain		0.0000	**0.0715**	**0.0591**	**0.0643**
20th Mexico			0.0000	0.0804	**0.0715**
18th Peninsula				0.0000	**0.0576**
20th Peninsula					0.0000

It is worth noting that the four smallest distances (numbers in bold) particularly occur between dialects of the 18th and 20th centuries. In fact, 20th century Mexican Spanish resembles 18th century Spanish in New Spain as much as it does 20th century Castilian. Furthermore, 16th century Spanish seems to be the most distant to all dialects (numbers in italics), as if a jump was made from that dialect to all other subsequent ones. However, the values are still very small. Again, one might think that the questionable representativity of such small samples would cause more noise, but the data seem to follow a pattern despite the small distances between dialects. Could it be that this exercise also attempts to discover patterns where none exist? The data in Table 10 can be summarized applying cluster analysis. Figure 2 shows results of hierarchical clustering.

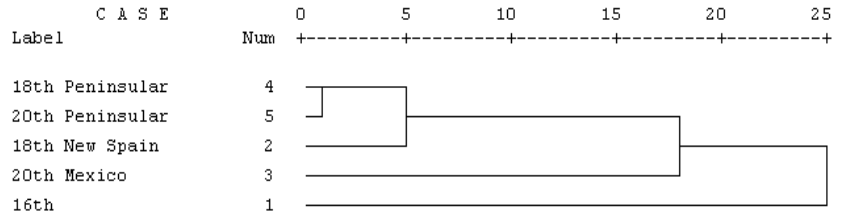


Figure 2. Dendrogram showing hierarchical clustering (nearest neighbor method, complete linkage) of Euclidean distances. These are rescaled in the range [0–25].

In this figure, the 18th and 20th centuries Peninsular dialects appear closest. The next similar profile is the New Spain's 18th century register. 20th century Mexican Spanish comes next. This set is next related to the 16th century profile. The fact that this latter dialect is the most morphologically distant to all other registers perhaps points out to a general change in the Spanish language in both sides of the Atlantic when the first wave of Europeans arrived in America. In short, the data seems to be consistent with what one would expect.

5. Concluding remarks

In this paper, quantitative data for three centuries of a dialectal subsystem of the Spanish language have been presented and contextualized. Data was extracted without supervision from samples of Spanish used in 16th, 18th and 20th centuries in what is today Mexico and the Spanish Peninsula. Although these samples are small and not really representative (except in the case of the CEMC), some intuitions concerning the morphological level can be preliminarily corroborated. For one, Mexican Spanish seems to have emerged as a dialectal system prior to the 18th century. Also, according to the data, contemporary Peninsular Spanish seems to be closer to 18th century Spanish of New Spain than to 20th century Mexican Spanish. This may very possibly refer to a bias in the written register. In any case, the conservative nature of the Spanish dialects seems to be corroborated.

Here I briefly described a simple method for estimating the affixality of strings (a type of glutinosity) based on the average of normalized square counts, entropy, and economy values. This method can be applied to compile unsupervised catalogs of affixes and affix groups for concatenative languages such as Spanish. In fact, these catalogs or morphological profiles appear to intimately characterize at least the dialects examined in this experiment. In this sense, they appear to represent true fingerprints. Besides this, simple Euclidean distances were calculated in order to measure distances between these profiles. The results presented are interesting, although hardly statistically significant. Nevertheless, measuring distances and similarities between sets of affixes and their sequences would appear to permit the comparison of diachronic stages within relatively short periods of time, at least in the case of written Spanish.

Undoubtedly, these experiments could be improved in many ways, for example, by testing other segmentation techniques, in order to measure affixality, taking into account glutinosity values for clitics and other modifiers, and applying alternative methods in order to measure distances and similarities between profiles. Finally, it would be valuable to apply these methods to other languages' dialectal systems in their diachronic and synchronic (geographic and social) dimensions. Is it possible to corroborate that the middle class register of the Spanish-speaking world represents a cluster of dialects with greater proximity between them than to any lower class register anywhere? How different are Chuj and Tojolabal, two Mayan languages, in their affixal morphologies? Could outstanding changes of 16th and 19th centuries Purépecha be examined by measuring Euclidean distances between their morphological profiles?

References

Creutz, M.and K. Lagus (2005), "Inducing the morphological lexicon of a natural language from unannotated text," in: T. Honkela, V. Könönen, M. Pöllä, and O. Simula (eds.) *Proceedings of AKRR'05, international and*

interdisciplinary conference on adaptive knowledge representation and reasoning, 106-113. Online at: <http://www.cis.hut.fi/mcreutz/papers/Creutz05akrr.pdf>.

Embleton, S. M. (1986), *Statistics in historical linguistics*. Bochum: Studienverlag Dr. N. Brockemeyer.

Goldsmith, J. (2001), "Unsupervised learning of the morphology of a natural language," *Computational linguistics*, 27(2): 153–198.

Greenberg, J. (1957), *Essays in linguistics*. Chicago: The University of Chicago Press, 1967.

Harris, Z. S. (1955), "From phoneme to morpheme," *Language*, 31(2): 190–222.

Kageura, K. (1999), "Bigram statistics revisited: a comparative examination of some statistical measures in morphological analysis of Japanese Kanji sequences," *Journal of quantitative linguistics*, 6: 149–166.

Kock, J. de, and W. Bossaert (1978), *The morpheme: an experiment in quantitative and computational linguistics*. Amsterdam/Madrid: Van Gorcum.

Lara, L. F. (1974), "Caracterización metódica del corpus del diccionario del español de México," in: L. F. Lara, R. Ham, and M. I. García (eds.) *Investigaciones lingüísticas en lexicografía*. Mexico: El Colegio de México, 5–39.

Medina Urrea, A. (2000), "Automatic discovery of affixes by means of a corpus: A catalog of Spanish affixes," *Journal of quantitative linguistics*, 7: 97–114.

Medina Urrea, A. (2003), "Investigación cuantitativa de afijos y clíticos del español de México: glutinometría en el corpus del español mexicano contemporáneo," Doctoral dissertation, Centro de Estudios Lingüísticos y Literarios, El Colegio de México, Mexico.

Medina Urrea, A., and E. C. Buenrostro Díaz (2003), "Características cuantitativas de la flexión verbal del chuj," *Estudios de lingüística aplicada*, 38: 15–31.

Medina Urrea, A., and J. Hlaváčová (2005), "Automatic recognition of Czech derivational prefixes," in: A. Gelbukh (ed.) *Computational linguistics and intelligent text*. Berlin: Springer, 184–192.

Medina Urrea, A., and M. Alvarado García (2006), "Un experimento de reconocimiento automático de la derivación léxica en el ralámuli," in: *La lengua y la antropología para un conocimiento global del hombre*, Mexico: Conaculta, INAH.

Medina Urrea, A. (2007), "Affix discovery by means of corpora: experiments for Spanish, Czech, Ralámuli, and Chuj," in: A. Mehler and R. Köhler (eds.) *Aspects of automatic text analysis*. Berlin: Springer, 275–297.

Oakes, M. P. (1998), *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Reyes Careaga, T. A. (2008), "Reglas de correspondencia entre sonido y grafía en el español hablado en México en el siglo XVI para la creación de un

transcriptor automático. Una aportación al CHEM," Thesis, Facultad de Filosofía y Letras, UNAM, Mexico.

Sapir, E. (1921), *Language: an introduction to the study of speech*. New York: Harcourt.

Shannon, C. E., and W. Weaver (1949), *The mathematical theory of communication*. Urbana: University of Illinois Press.

Swadesh, Morris. (1955), "Towards greater accuracy in lexicostatistic dating," *International journal of American linguistics*, 21: 121-137.

Change and variation in complement selection: a case study from recent English, with evidence from large corpora

*Juhani Rudanko**

University of Tampere

Abstract

The study examines change and variation in the system of English predicate complementation in recent times on the basis of three major corpora, the Corpus of English Novels (the CEN), comprising some 18 million words, the full Bank of English Corpus, comprising some 524 million words, and the Corpus of Contemporary American English, comprising some 360 million words in the version used, with the focus on the matrix predicate submit *and its sentential complements. The sentential complements in question are of the* to *infinitive and* to -ing *types. It is shown that there are sharp grammatical differences between these two types of complement in English. However,* submit *is a matrix verb that has selected both types in recent English. In the CEN, representing usage from about a century ago,* to *infinitive complements predominated over* to -ing *complements by a ratio of over five to one, but in current English,* to -ing *complements predominate over* to *infinitives by a ratio of almost two to one. The study examines the nature of the variation and change, and it is pointed out that, in spite of the sharp grammatical differences between the two patterns, in the CEN both types of complement were found in one and the same text by one author, without any apparent distinction relating to genre, register, or style between the two variants. A difference in meaning is also hard to establish on the basis of contrasts generally posited in the literature for* to *infinitives and* -ing *forms. However, the study shows that passive and passive-like lower predicates with lower subjects of the type Patient or Undergoer are often associated with* submit*, and it is suggested that this association may have promoted the grammatical change from* to *infinitives to* to -ing *complements in the present case, in the overall context of what has come to be called the Great Complement Shift in the recent literature. The* -ing *pattern is virtually unique to English as a pattern of nonfinite complementation, and the grammatical change in question, it is suggested, is an example of system-internal change. The present study suggests that it is desirable to conduct follow-up work on the nature of the semantic roles of lower subjects as a factor bearing on grammatical change in the case of predicates selecting nonfinite sentential complements in recent centuries.*

1. Introduction

Consider (1a–b):

(1) a. John wanted to change the agreement.
b. John objected to changing the agreement.

Sentences (1a) and (1b) share a number of properties. In both the word *to* is found, and both involve a sentential complement. The pattern of (1a) is here termed the *to* infinitive pattern, and the pattern of (1b) is termed the *to -ing* pattern. In the latter the *-ing* clause might also be called a gerund. It is also assumed here that in each case the lower clause has its own subject. This assumption is somewhat controversial, but it was made by many traditional grammarians, including Otto Jespersen (1961 [1940]: 140). He made this comment on gerundial *-ing* clauses:

> Very often a gerund stands alone without any subject, but as in other nexuses (nexus substantives, infinitives, etc.) the connexion of a subject with the verbal idea is always implied.

The assumption of an understood subject in sentences of the type of (1a–b) is also made in more recent work. Chomsky (1986: 114–31) makes it, and so do Huddleston and Pullum (2002: 1193) in their grammar. Apart from an appeal to tradition, it may also be observed that the postulation of an understood subject makes it easy to represent the argument structures of the verbs in (1a–b), with the understood subject representing the subject argument of *change*. In line with current work, the understood subject is represented here by the symbol *PRO*. However, there are also robust grammatical differences between the two patterns. First of all, the patterns are clearly distinct in that they are not interchangeable, at least not with the higher verbs *want* and *object*:

(2) a. *John wanted to changing the agreement.
b. *John objected to change the agreement.

Second, an NP complement can follow the word *to* in (1b), but not in (1a):

(3) a. *John wanted to change the agreement, but Mac did not want to it.
b. John objected to changing the agreement, but Mac did not object to it.

By contrast, VP Deletion is possible in the case of (1a), but not in the case of (1b):

(4) a. John wanted to change the agreement, but Mac did not want to.
b. *John objected to changing the agreement, but Mac did not object to.

To explain such differences as those observed in (3a–b) and (4a–b), it is proposed here, in line with Quirk et al. and other work, that there are two types of *to* in present-day English. The *to* of (1a) is an infinitive marker (Quirk et al., 1985: 1178, note a) or an Aux or an Infl (Chomsky, 1981: 18 f.). By contrast, the word *to* of (1b) is a preposition. It is also assumed that the complement of (1b) is a nominal clause, that is, a sentence dominated by an NP node. The structures in question are given in (1a´) and (1b´).

(1) a´. [[John]$_{NP1}$ wanted [[PRO]$_{NP2}$ [to]$_{Infl}$ [change the agreement]$_{VP}$]$_{S2}$]$_{S1}$
b´. [[John]$_{NP1}$ objected [[to]$_{Prep}$ [[PRO]$_{NP2}$ changing the agreement]$_{S2}$]$_{NP}$]$_{S1}$

VP deletion can apply in (1a´) since the constituent that follows the word *to* is a VP. This is not the case in (1b´). By contrast, in (1b´) the constituent that follows the word *to* is an NP, which explains why *it* and *that*, which are pro-forms for NPs, are possible in this case.

In spite of the sharp grammatical differences between the two patterns, it has been observed in some earlier work, including Poutsma (MS), Kjellmer (1980), Denison (1998), and Rudanko (1998a, 1998b), that there are a number of matrix predicates—verbs, adjectives, nouns—that have shown variation and change between the two patterns in recent times. The purpose of this study is to examine one predicate of this type, the verb *submit*.

The second edition of the OED contains a comment on what are here termed *to* infinitive and *to -ing* complements of *submit*. The comment comes under sense two of the verb, which is formulated as follows:

> 2. To surrender oneself *to* judgment, criticism, correction, a condition, treatment, etc.; to consent to undergo or abide by a condition, etc.

Illustrations of the sense are divided into reflexive and nonreflexive usages, with the latter termed "*intr*.," as in (5a) and (5b), respectively:

(5) a. The majority of cases would voluntarily submit themselves to treatment. (1913, OED, *Times*)
b. Healing measures … such as … all men must, with more or less reluctance, submit to. (1837, OED, Carlyle)

It may be observed how in both (5a) and (5b) the complement of the verb is of the *to* NP type. The comment on *to* infinitive and *to -ing* complements is supplied under sense 2.b of the verb:

> 2.b Const *to* with inf. or gerund: To yield so far as *to do* so-and-so, consent *to*; *occas*. to condescend *to*. *Obs*.

Several illustrations are offered of the usage in the OED, and these are again grouped under the reflexive and intransitive labels. The former group contains only archaic tokens, from pre-Shakespearean times, but in the latter group there are tokens from recent centuries. Here are the four illustrations in the OED that are the most recent:

(6) a. They, at last, submitted, to have these words left out. (1697, C. Leslie)
b. She submitted to humble herself to Montoni. (1794, Mrs. Radcliffe)
c. Where the mortgagee submits to be redeemed. (1818, Cruise)

d. I … affected gladness when he came, submitted to hear when he was by me. (1852, Thackeray)

As regards the OED treatment of the sentential pattern being investigated, two points deserve to be singled out for attention. First, the OED mentions *to* infinitival and *to -ing* (or gerund) clauses side by side as complements with sense 2b. However, in illustrations complements of the *to* infinitive type predominate, as in the most recent ones in (6a–d), and there do not appear to be any illustrations of *to -ing* complements in the OED treatment of the verb.

The other point is that even though illustrations of sentential complements are supplied from recent centuries, including one from as late as Thackeray, both types of usage are claimed to be obsolete in the second edition of the OED, which is from 1989. It may be added that the comment that both types of usage involving sentential complements are obsolete is repeated in the electronic version of the OED as of February 28, 2008.

For his part, the great Dutch grammarian H. Poutsma offered a number of illustrations of the verb with sentential complements involving subject control, including (7a–c).

(7) a. He submitted to be kissed willingly enough. (G. Eliot, *Mill*, Ch. V)
b. Could he ever submit to give up Sibyl for any other? (Mrs. Alex., *For his Sake*, Ch. II)
c. It is possible that the population of the Ruhr may submit to working for the French. (*Manch. Guard. W.*)

A problem with Poutsma's illustrations is the absence of dates, but many of them come from major authors, as in (7a–b), and in their case the problem is not serious. As for (7c), the content may presumably help date the example as later than 1920. Several of Poutsma's illustrations have *to* infinitive complements, as in (7a–b), but he also offered an illustration of one *to -ing* complement, reproduced in (7c).

> He then offered this comment on the OED treatment of the verb:
> The O.E.D. calls the use of the gerund- and the infinitive-construction obsolete. But this view can hardly be upheld, so far, at least, as the latter is concerned, which appears to be quite common, especially with this verbal in the passive voice. (Poutsma, MS)

The purpose of this study is to examine both types of complement in usage about a century ago, in the period 1880–1922, and in current English on the basis of electronic corpora. The diachronic corpus examined is the Corpus of English Novels (the CEN), compiled at the Catholic University of Leuven, with Hendrik de Smet as the main originator of the corpus. It has material from both British and American English, and for present purposes, two parts are used, a British English part, of some 12.3 million words, and an American English part, of some 5.9

million words. For present-day English, the Bank of English Corpus, of some 524 million words, and the newly released Corpus of Contemporary American English are examined.

2. Sentential Complements of *submit* in the CEN

Turning to the British English segment of the CEN, it must be pointed out that the corpus is untagged and no tags can be used in search strings. The four simple search strings *submit*, *submitted*, *submits*, and *submitting* were therefore used. These search strings can be expected to give total or near-total recall.

Table 1 gives the number of hits for each search term:

Table 1. Hits in the British English segment of the CEN.

submit	173
submitted	133
submits	4
submitting	29

However, the number of relevant tokens is only a fraction of the total of more than 300 hits. Irrelevant tokens are of different types. First, reflexive uses of *submit*, as in (8), are set aside:

(8) … on the whole it seemed probable that Louise would conscientiously submit herself to instruction.… (1895, George Gissing, *The Paying Guest*)

Among nonreflexive uses there are other syntactic patterns, where the sense of *submit* 'to surrender oneself to judgment, criticism, a condition, treatment, etc.,' quoted above, is not relevant. For instance, there are tokens of *submit* with NP *to* NP complements, as in (9):

(9) Then I'll just say that I submit to them a novel of modern life, the scope of which.… (1891, George Gissing, *New Grub Street*)

At the same time, the 'surrender oneself to something' sense of *submit* is a common sense in the material. This sense is very frequent with *to* NP complements, as in (10a–b):

(10) a. She did indeed feel ill, but to submit to treatment was impossible whilst this day lasted. (1888, George Gissing, *A Life's Morning*)
b. Perhaps she was moved to this merely by a desire to submit to her husband's will, and to realise his hopes and expectations. (1913, Hall Caine, *The Woman Thou Gavest Me*)

The present investigator has counted 64 tokens of this pattern among the 173 tokens of *submit* in the material. The same sense is also frequent in

constructions where the complement is omitted, with 45 tokens in the case of the verb form *submit*. Two illustrations are given in (11a–b):

(11) a. She controlled herself and began to write. There was no escape. She must submit; and all was over. (1903, Humphry Ward, *Lady Rose's Daughter*)
b. … who has already promised pardon to all Poles who have taken arms against Russia and now submit. (1903, Henry Seton Merriman, *Barlasch of the Guard*)

Sentential complements are less frequent, but there are 27 of them in the material. This represents a frequency of 2.2 per million words. Twenty-three of these are of the *to* infinitive type, and four are of the *to -ing* type, representing frequencies of 1.9 and 0.3 per million words, respectively. That is, *to* infinitives dominate over *to -ing* complements by a ratio of more than five to one. Here are all four tokens of the *to -ing* type.

(12) a. … when Betty first demanded to know what she was going to wear, and then pouted over the dress shown her, Marcella submitted humbly to being "freshened up" at the hands of Lady Ermyntrude's maid,… (1894, Humphry Ward, *Marcella*)
b. In time he submitted even to being flown at and kissed before the Fullertons. (1896, Humphry Ward, *Sir George Tressady*)
c. Perforce she submitted to having her hair done by her maid, but she found the necessity disagreeable. (1897, George Gissing, *The Whirlpool*)
d. Why had she been such a fool as to come to Monk Lawrence at all, and then to submit to seeing it—on sufferance!—in Winnington's custody. (1914, Humphry Ward, *Delia Blanchflower*)

It is observed how three of the four tokens of *to -ing* complements come from one author, Humphry Ward, with the third coming from George Gissing. Under these circumstances, it is of interest to see if tokens of *to* infinitive complements can be found in the works of these same authors. The answer is yes.

(13) a. The miserable wife submitted to be fed, looked with forlorn wonder at the children round the fire, then sank back with a groan. (1894, Humphry Ward, *Marcella*)
b. While he was speaking she had a slight return of pain, and was obliged to submit to lie down again. (1896, Humphry Ward, *Sir George Tressady*)
c. … he had seen a priest submit to be dragged on his back across a turnip field,… (1899, George Gissing, *The Crown of Life*)
d. He had soon reconquered cheerfulness; and when Arthur returned, he submitted to be talked to for hours on that young man's tangled affairs,… (1913, Humphry Ward, *The Coryston Family*)

It is observed how *submit* selects both *to* infinitive and *to -ing* complements in one and the same book of an author, as in the case of (12a–b) and (13a–b), or in books by the same author that are almost co-contemporaneous, as in the case of (12c–d) and (13c–d). Given the data in (12a–d) and (13a–d), an obvious question to ask is whether or not it is possible to point to a semantic difference between the two types. Here we might recall David Allerton's (1988) list of factors impacting the choice between infinitives and gerunds, one of the fullest lists in the literature. As he puts it, the "infinitive-gerund distinction, in its healthy state, can be summed up with the following features" (Allerton 1988, 21):

Table 2. Allerton's summary of the infinitive-gerund distinction.

Infinitive	**Gerund**
infrequent activity	regular activity
intermittent activity	continuous activity
interrupted activity	continuing activity
uncompleted activity	completed activity
contingent/possible event	event presented factually
particular time and place	neutral time and place
specific subject	nonspecific subject
more verbal character	more nominal character

Notions similar to those in Allerton's summary are found in many other scholarly attempts in the literature to distinguish infinitives and *-ing* forms, but it is not easy to apply such notions in the present case. For instance, with respect to the first distinction between "infrequent activity" versus "regular activity," the notion of a regular activity may be relevant to (12b), but even though (12a) has a *to -ing* complement, the scenario involved, with *first … then*, suggests a one-off event rather than a regular activity. For their part, several of the *to* infinitive complements in (13a–d), including (13a), convey a completed activity rather than something uncompleted. Further, with both infinitival and *to -ing* complements, the subject of the lower clause is coreferential with the higher subject in all the tokens, and the specific versus nonspecific subject distinction therefore does not seem applicable in the present case.

Overall, it seems difficult to identify a consistent and stable semantic difference between the complements in (12a–d) and those in (13a–d). While subsequent research may yet identify such a difference, it may be that it will not, given the similarity of the complements. If a stable semantic difference cannot be made out, it is appropriate—recalling the sharply differing grammatical properties of the two types of complement—to refer to the availability of competing grammars for both Humphry Ward and for George Gissing on the basis of the argument structure properties of *submit*. Or, more precisely, the competition is between two distinct alternatives or variants within a grammar. This point is made succinctly by Pintzuk (2003: 516) when she observes that "it is not two entire grammars that are in competition, but rather two contradictory options within a grammar."

Competing grammars or the competition between two contradictory options within a grammar are highly salient to grammatical change because this type of variation is "diachronically unstable" (Pintzuk, 2003: 510). Regarding the nature of the competing forms, Kroch (2001: 702–703) writes:

> The best-studied cases of long-term syntactic drift are most plausibly cases of grammar competition (that is, syntactic diglossia) in which the competing forms may differ in social register, with an unreflecting vernacular variant slowly driving a conservative written one out of use.… Where no such process is at work, there is evidence that usage frequencies remain stable over long periods of time.

If the notion of syntactic diaglossia is understood in this way to refer to variants that belong to different social registers, with one grammar or one variant within a grammar used "for formal, institutional contexts and the other in less formal contexts" (Roberts 2007: 324 f.), the present case is of some interest. For instance, we may compare (12a) and (13a). The sentential complements of *submit* are sharply different from a grammatical point of view, but it is hardly possible to discern any difference in terms of register separating the two uses of *submit*. In both cases, the complements are found in the narration of the author, and they are as similar as can be with respect to social register.

While the numbers of *to -ing* complements are low and additional work on the two patterns of sentential complementation during the period 1880–1922 is desirable when even larger corpora become available, the findings relating to the usage of Humphry Ward and George Gissing suggest that, in the absence of a clearly definable semantic difference between the two patterns, it may be necessary to postulate non-diaglossic multiple options within grammars for individual speakers during a period of grammatical change.

One perspective that may be relevant to understanding the nature of sentential complements with *submit* and a reason for change in this case is the nature of the lower sentence. It is conspicuous that of the 27 tokens, as many as 21 involve a passive complement. Illustrations include (12a) and (13a), among others. In such sentences, the semantic role of the lower subject is typically of the type of Patient or Undergoer. Table 3 gives information about the pattern from this point of view:

Table 3. Active and passive lower clauses with *submit* (BrE part of CEN).

	***to* infinitive**	***to -ing***
passive lower clause	18	2
active lower clause	5	2

The *to* infinitive pattern clearly dominates with passive lower clauses, but the low numbers of *to -ing* complements make it difficult to make any claim of statistical significance about variation between the two patterns involving lower clauses in the passive in the material. However, what is noteworthy is the high

frequency of passive lower clauses, and because of their preponderance, it may be possible to view the passive pattern as a specific construction in the case of *submit*.

In the American English part of the CEN, there are 13 tokens of sentential complements of *submit*, which is 2.2 tokens per million words. This is much the same frequency as in the British English data. The number of *to* infinitive complements among the 13 is 9, representing a frequency of 1.5 per million words, and the number of *to -ing* complements is then 4, representing a frequency of 0.7 per million words. While the numbers are low, the proportion of *to -ing* complements is thus much higher in the American English data than in the British English data. Here are illustrations:

(14) a. False as was their accusation, she submitted to hearing her father speak them, for she had no knowledge of their import,… (1884, Francis Marion Crawford, *A Roman Singer*)
b. … it was impossible that any misunderstanding should last long, for he was too honest and frank to submit to being misunderstood himself. (1889, Francis Marion Crawford, *Greifenstein*)
c. "Sir, you are in grave danger; we are both in grave danger," he announced, "unless we choose to submit to being robbed by this rascally brigand." (1906, Lyman Frank Baum, *Aunt Jane's Nieces Abroad*)
d. Undine, after this, submitted in brooding silence to having her dress unlaced,… (1913, Edith Wharton, *The Custom of the Country*)

(15) a. But Greif would not submit to be treated like a child, and sprang up, seizing the man's arm and drawing him nearer. (1889, Francis Marion Crawford, *Greifenstein*)
b. She was wondering why she had submitted to be betrothed to Contarini, when she loved Zorzi; and the answer did not come. (1901, Francis Marion Crawford, *Marietta*)
c. "I have a cab," replied Donna Tullia, faintly, submitting to be put out of the door. (1887, Francis Marion Crawford, *Saracinesca*)
d. The remuda was young, gentle, and sound, many of them submitting to be caught without a rope. (1904, Andy Adams, *A Texas Matchmaker*)

The last two illustrations suggest a *horror aequi* effect (Rohdenburg, 2006). There are four tokens of the verb form *submitting*, and in all these cases, the complement is of the *to* infinitive type, in accordance with the *horror aequi* principle. The numbers are too low for statistical significance, but are nevertheless worth noting. (14b) may likewise involve a *horror aequi* effect, with *to submit* followed by *to being*, but it is still worth noting that both types of complement are found in Francis Marion Crawford's *Greifenstein*. Information about the form of the lower clause in the American English material is given in Table 4.

Table 4. Active and passive lower clauses with *submit* (AmE part of CEN).

	***to* infinitives**	***to -ing***
passive lower clause	8	3
active lower clause	1	1

Once again, the numbers of tokens are so low that it is difficult to make any claim of statistical significance about variation between *to* infinitives and *to -ing* complements with lower clauses in the passive. However, what is of interest is that the American English data are consistent with the finding above that the lower clause is typically in the passive in sentential complements of *submit*.

The evidence of the CEN shows that sentential complements were found with *submit* in both British and American English in the period 1880–1922. It further shows that the verb selected both *to* infinitive complements and *to -ing* complements involving subject control in this period. As argued above, in the former the *to* is an Aux, and in the latter it is a preposition.

Overall, *to* infinitives were considerably more frequent than *to -ing* complements, and the predominance of the former was much more pronounced in British English than in American English: 22 versus 4 and 9 versus 4.

3. Sentential complements of *submit* in the Bank of English Corpus

Turning to present-day English, the full Bank of English Corpus is a suitable resource for the study of the verb. The corpus was investigated in February 2008. The original plan was to choose two or three British English and two or three American English corpora in order to study the sentential complements in current English. For this purpose, a pilot study was made of the pattern "*submit*@" in the *Sun* and *News of the World* (SUNNOW) corpus of the Bank of English Corpus. This corpus is large, some 45 million words. The search with the search string produced 240 hits. Among them, there are 74 tokens of the verb form *submit* and 4 of *submits*. Among these tokens, there are numerous tokens of the pattern *submit* NP (*a complaint, a report*, etc.) *to somebody*, which can be set aside as irrelevant, but the relevant sense of 'surrender oneself to something' is also found. However, it is only found with *to* NP complements and in sentences lacking complements, as in (16a–b), respectively.

(16) a. Would you lie back and submit to [a name]'s bedside manner? (SUNNOW)
b. And [a name] refuses to rule out doing more raunchy nude photo shoots. But for a girl who submits so sweetly she's not above laying down the law. (SUNNOW)

Because of the results of the pilot study, which failed to yield any relevant sentential complements, the decision was made not to choose subcorpora of the Bank of English Corpus for this study and instead to use the entire Bank of

English Corpus, of some 524 million words. The decision was also made to use a specific set of search strings. These are given in (17).

(17) submit@+to+VBI
submit@+to+VBG
submit@+to+be
submit@+to+being
submit@+to+have
submit@+to+having

Submit@ is a search expression that retrieves all the different forms of the verb *submit*. "VBI" is the tag for an infinitival verb form, and "VBG" is the tag for an *-ing* form. However, the auxiliary verbs *be*, *have* and *do* have special tags for their infinitive and *-ing* forms, and therefore additional search strings were added. There are only four of these in (17), because no tokens were found of the type *submit@+to+do* or of *submit@+to+doing*. These two strings are therefore not included in (17).

From the point of view of recall, the six search strings may be expected to be comprehensive, covering all forms of the verb *submit*. Regarding the first two search strings, they are based on tags. Errors in labeling may cause some tokens to be omitted here, in cases where a verb has been mislabeled as something else, but this caveat attaches to virtually any search string making use of tags, and the large majority of relevant tokens can be expected to be retrieved by the tags.

The search strings are specific, but in the case of "*submit@+to*+VBI," quite a number of irrelevant tokens were encountered. These are of several different types. Two are illustrated in (18a–b):

(18) a. I mean, I have an item that I'll submit to replace this, … (USSpok)
b. Why would a president submit to press conferences when he can chat up Larry King for an hour or so every couple of weeks and speak directly to the public? (USSpok)

(18a) shows that it is important for the investigator to be sensitive to syntactic operations in sentences, and the sentence can be excluded as irrelevant here. As for (18b), it shows how elements in electronic corpora are sometimes tagged incorrectly, and it is easy to exclude tokens of the type of (18b). In addition, there are nine tokens that are much less easy to decide on. Consider (19a–b):

(19) a. President Bush refused to say today whether he would submit to questioning next month as requested by Walsh. (USSpok)
b. … [a name] had learned only last week that "a major player submitted to interviewing by government agents." (USNews)

The question is whether the *-ing* forms in (19a–b) are sentential or nominalizations. These are borderline cases, and a decision is made more difficult by the fact that *-ing* complements, even when sentential, are at the nominal end of sentential complements. However, while *-ing* forms of the type of *questioning* and *interviewing* are worth noting, it seems appropriate to exclude them here as nominal *-ing* forms. One criterion for deciding on the status of an unclear *-ing* complement that goes back to Wasow and Roeper (1972) is whether an adjective can be inserted in front of the *-ing* complement: if an insertion is possible, this suggests that the *-ing* form is nominal. Such an insertion seems possible in the present case, though admittedly it is more natural in the case of *questioning* than in the case of *interviewing*, as in … *he would [not] submit to hostile questioning*.

In order to err on the side of caution in the consideration of *to -ing* complements and in order not to exaggerate their role with *submit*, the cases of (19a–b) are excluded here as nominal, though it should be noted that these are borderline decisions, especially in the case of *interviewing*. With the exclusions carried out, the findings for the different search strings are as in Table 5.

Table 5. The numbers of tokens for each search string in the Bank of English Corpus.

Search string	**Tokens**
submit@+*to*+VBI	3
submit@+*to*+VBG	2
submit@+*to*+*be*	7
submit@+*to*+*being*	11
submit@+*to*+*have*	2
submit@+*to*+*having*	7

Here are illustrations of the three pairs:

(20) a. … they could not understand why his Swiss guards, so many grown men, bearded, strong, and armed … should submit to obey a child instead of putting one of their members in command. (US Books; the three periods in the middle as in the original, JR)
b. "… if it will save him an hours disquiet I readily submit to continue his steward." (Strathy)
c. … Although "as the wife of a county practitioner" Mary would have to "submit to associate with vulgarity, even more conspicuous than that which.…" (Strathy)
d. Under protest, Rendell had submitted to giving a deposition in a federal court lawsuit stemming from a 1998 incident outside City Hall.… (USNews)
e. … few women have more willingly submitted to becoming the passive material out of which a myth can be created.… (BrNews)

(21) a. Perhaps he was trying to seek salvation and Annabella saw herself as his redeemer. Yet the misery of their year-long marriage showed that Byron

could not submit to be tamed and Annabella had shackled herself to a powerful and destructive force. (BrNews)

b. If a single bedroom can be procured, more ought not to be looked for; but it is not always that even this is to be had, and those who travel through the country must often submit to be crammed into rooms where there is scarcely sufficient space to walk between the beds. (US Books)

c. We were visited by a small party of Indian entertainers, including a young man who claimed to be a fakir (he was also a university graduate). He was certainly highly intelligent. He submitted to being buried for about 25 minutes, several feet deep, with only a square of sackcloth over his face. (The *Times*)

d. But opponents say Web marketers such as doubleclick should track only consumers who opt in, or voluntarily submit to being profiled. (BrMags)

(22) a. In order that Mr. Sifton may keep his Liberal party in power by the votes of ignorant and vicious foreign scum he is dumping on our prairies, we are to submit to have our nearest and dearest butchered on our door-steps. (Strathy)

b. … the Defendant should have an opportunity of choosing whether he will submit … to have the lease rectified … or … choose to throw up the thing entirely.… (Strathy; the three periods in the middle of the extract are as in the original in each case, JR)

c. I mean, you didn't have to eat your vegetables once you got out of the family household, and why should you submit to having your friends tell you what you have to eat and drink? (USSpok)

e. Meekly each of the family members submitted to having the caustic red poison painted gaggingly into their throats. (USBooks)

The overall number of *to* infinitive complements is 12, and that of *to -ing* complements is 20. *To -ing* complements are clearly predominant in relation to *to* infinitive complements in current English in sentential complements of *submit*, but there are still sizeable numbers of *to* infinitives found as well.

The material shows that *to* infinitive complements are found with *submit* in current English, but their proportion is clearly much lower in relation to *to -ing* complements than what was the case a century ago. There is still variation between the two complements in current English, but it is possible to say on the basis of the present material that the *to -ing* pattern has gained ground at the expense of the *to* infinitive. The change is part of a broader set of changes in the system of English predicate complementation, and the term Great Complement Shift has been used in the literature to highlight the amount of change in question. The increasing prominence of *-ing* complements, whether straight, that is, non-prepositional, or prepositional, as in the present case, is an important part of the Shift (see Rohdenburg, 2006).

To explain the increasing prominence of the *to -ing* pattern in relation to the *to* infinitive pattern, a number of considerations may be adduced. Denison (1998) examined the matrix verb *object* and some other verbs that show this

development affecting sentential complements and suggested, firstly, that *object* and *to* have come to form a close association and, secondly, that the *to* infinitive has undergone a long-term trend "from a nominal to a verbal character, now virtually complete" and that this trend has also resulted in the "concomitant dissociation of the infinitive marker *to* from the homonymous preposition" (Denison, 1998: 266).

Denison's study was published in 1998. In the same year, the present author published two studies that also highlighted the spread of the *to -ing* pattern at the expense of *to* infinitive complements, and one point made in this work was that *to -ing* complements were—and are—more likely to emerge and to spread in the case of higher predicates—verbs, adjectives, nouns—that also select *to* NP complements (Rudanko, 1998a; 1998b). As was emphasized in John Robert Ross's important work in the 1970s (for instance, Ross, 1973), *-ing* complements are more nominal, or nouny, in nature than *to* infinitive complements, as also pointed out by Denison. It therefore stands to reason that *to -ing* complements should be found where non-sentential *to* NP complements are found. It is recalled how *to* NP complements were very common with *submit* in the CEN, and the stage was thus set for the spread of the *to -ing* pattern.

What may also have facilitated the general spread of *to -ing* complements in the case of *submit* is the association of the pattern with passive lower clauses. Such clauses, readily compatible with the 'submissive' sense of *submit*, were predominant in the diachronic corpora, as noted above, and they are similarly predominant in the Bank of English Corpus: of the 32 tokens of sentential complements, 15 are in the active and 17 are in the passive. These figures are based on the form of the verb directly embedded under *submit*, and additionally, there are 6 tokens of *submit to have/having* constructions that are passive in a broader sense in that in them the verb in the subordinate sentence immediately below *have* is passive, as in (23):

(23) Martine submits to having her shoes put on for her.… (BrNews)

These six tokens are also passive in nature, and the semantic roles of the lower subjects of such sentences are also of the type Patient or Undergoer. When these tokens are moved to the passive column, the figures change to 9 in the active and 23 in the passive. It has been argued in earlier work that passive and passive-like lower complements are especially compatible with *to -ing* complements (Rudanko, 2006), and the attraction of *submit* to passive lower clauses may likewise have contributed to the spread of *to -ing* complements with this verb.

The evidence of the Bank of English Corpus testifies to a general spread of the *to -ing* pattern at the expense of the *to* infinitive pattern in complements of *submit*, but the corpus also affords an opportunity to examine the spread in relation to regional variation. The total of sentential complements in the entire corpus is 32, and Table 6 gives information on the regional distribution of the two different types of sentential complements. The corpus size is given in millions of

words and the frequencies have been calculated per million words. The figures in parentheses likewise indicate frequencies per million words.

Table 6. Frequencies of *to* and *to –ing* in the regional subcorpora of the *Bank of English*.

	To	***To -ing***	**Total**	**Corpus size**
BrE	4 (0.01)	13 (0.04)	17	346.8
AmE	2 (0.02)	5 (0.04)	7	126.5
Canadian	6 (0.39)	2 (0.13)	8	15.5
Oceanic	0	0	0	26.1
Total	12 (0.02)	20 (0.04)	32	524.3

The numbers of tokens, unfortunately, are rather small, in spite of the large size of the Bank of English Corpus, and they should be viewed with caution. However, three findings do emerge, even if they are in part in the nature of invitations to carry out further work rather than final results.

The first is that sentential complements of *submit* have become much more rare in current English than they were a century ago. It is recalled that in the British English part of the CEN, the frequency of such sentential complements was 2.1 per million, but in the present material, it is only about one twentieth of that. In the American English part of the CEN the frequency was 2.2 per million words, and now it is again only about one twentieth of that. Naturally, there may be some effect from text type, and it might be that sentential complements are more common in imaginative fiction than in other text types. However, only a handful of the present-day English tokens of sentential complements come from the text type of books, British or American, even though the Corpus contains two subcorpora of books of over 50 million words each. It does appear, therefore, that sentential complements have become much more rare with *submit*.

A second finding is that there is no dramatic difference between British and American English with respect to the two types of sentential complements. Both types are attested in both varieties, and in both *to -ing* complements are much more frequent.

Third, what does stand out from Table 4 is the relatively high proportion of sentential complements in Canadian English, and, further, that a high proportion of such complements is of the *to* infinitive type in Canadian English. At the same time, two caveats should be made. First, the overall numbers are low, and second, on closer inspection, it turns out that three of the six tokens found come from a single text. The finding about Canadian English is therefore an invitation to undertake further work on the construction in Canadian English, rather than a final and confirmed result.

4. Sentential complements of *submit* in the Corpus of Contemporary American English

The *Corpus of Contemporary American English* is a corpus of some 400 million words of recent American English. The corpus became available in February 2008, and it was accessed on February 21, 2008. When it was accessed, the corpus comprised some 360 million words. To investigate sentential complements of *submit* involving subject control, a search using the search string "[submit] to [v*]" was carried out. The first term designates a lemma search for the four forms of the verb *submit*, and the third term covers all verbs. The search string is specific, and 93 tokens were retrieved with it from the corpus. In spite of the specific nature of the search string, a large majority of the 93 tokens are irrelevant. Sometimes such irrelevant tokens result from occasional errors in tagging, as for instance in (24a–b).

(24) a. Thereafter they could go as they pleased but had to submit to brief checks concerning contagious diseases for another three days. (1993, ACAD)
b. … spend hours each day building our bodies, or submit to face lifts, breast implants, tummy tucks, liposuction. (1993, ACAD)

There are also fairly numerous tokens where what follows *to* is a verb but where the pattern is quite different from the subject control pattern dependent on *submit* under review here, as in (25a–b):

(25) a. These letters were submitted to be entered in the Congressional Record.… (1996, MAG)
b. The insurance company also monitors bills submitted to be sure that no one is exploiting the system. (2004, MAG)

The present investigator has been able to find 19 tokens of the pattern under review among the 93 tokens, which represents a frequency of 0.05 per million words. This is similar to the frequency of the pattern in the British English and American English segments of the Bank of English Corpus considered in Section 3. Of the 19 tokens, 4 are of the *to* infinitive type, and 15 are of the *to -ing* type. Here are the four *to* infinitives and some illustrations of *to -ing* complements:

(26) a. Furthermore, the militia would not be sufficiently trained because its members "would not long, if at all, submit to be dragged from their occupations and families." (1993, ACAD)
b. … this group "allegorically represents the power of beauty over savage nature. The monarch of the forest. Unable to resist the seducing loveliness of a nude female who is seated on his back and fascinating him with her eyes, is quietly submitting to be deprived of his claws." (1994, FIC)

c. … the final overthrow or expulsion of the Mound Builders was sudden and complete. It was so sudden that the mines of Lake Superior were abandoned in such haste as to cause them to leave their implements behind. On the temple mounds were probably scenes of carnage. They never would submit to give up these places without first offering the most stubborn resistance. (1996, ACAD)

d. The young lady he was addressing—she had the wondrous name of Augusta Tweddell—had sent him a hopeful packet of manuscripts, and advice came pouring out. "Will you submit to be preached at for a while?" (2003, ACAD)

(27) a. … there are certain sins for which a person must submit to being righteously slain. (1999, MAG)

b. It was bad enough to have to submit to getting a gamma globulin shot because.… (2004, FIC)

c. … when people submit to wearing their uniform they are necessarily obliged to another set of values and beliefs.… (2007, ACAD)

d. She even submitted to having her graceful wings unfolded to satiate our curiosity about her wingspan,… (1992, MAG)

e. The play is in your freely and autonomously submitting to being tied up. (1998, MAG)

The numbers of tokens are low, but the evidence of the Corpus of Contemporary American English suggests that both types of complements are still found with *submit*. However, it also shows that the proportion of *to* infinitive complements is clearly much lower than that of *to -ing* complements. As is seen in several of the illustrations, the lower clause is often in the passive, which is the case in at least 13 of the 19 tokens, which may have promoted the spread of *to -ing* complements.

When we look for a reason to explain the presence of *to* infinitives in (26a–d), it is observed that none of the *to* infinitives in (26a–d) involves extraction and that this factor does not appear to play a noteworthy role in current American English in the case of the two variants of *submit*. This suggestion is also supported by the presence of extraction in (27a), which has a *to -ing* complement rather than a *to* infinitive.[1] What seems relevant as a factor "protecting" *to* infinitives is simply the historical flavor of at least some of the tokens. (26a) is from a discussion of Alexander Hamilton, and the token may be a quotation from the period. (26c) appears to describe an event in the nineteenth century, and as for (26b), it appears to be from a discussion of Kipling.

Tokens of *to -ing* complements are clearly much more frequent in the new corpus than *to* infinitive complements, and from a qualitative point of view, it is observed how a *to -ing* complement is found in (27e) in spite of the context that might be expected to favor a *to* infinitive on account of the *horror aequi* principle. However, the overall frequency of *to -ing* complements is not high. Fifteen tokens in a corpus of 360 million words only amounts to a normalized frequency of 0.04 per million. The corresponding normalized frequency of the

pattern in the American English part of the Bank of English Corpus is about the same at 0.04 per million, and the new corpus thus further testifies to the rarity of even the *to -ing* pattern. As regards the nature of the 15 tokens, it is observed that passive lower clauses form a large majority among them, with 11 tokens. The number of active lower verbs is only four, but there is some variety among the verbs in question, as illustrated by (27b), (27c), and (27d).

5. Conclusion

The present study has focused attention on change and variation in the system of English predicate complementation in recent times on the basis of two major corpora of English, with the focus on the matrix verb *submit* and its sentential complements. Both *to* infinitive and *to -ing* complements can be found in constructions involving subject control, but it was argued that there exist sharp grammatical differences between the two types of complement. In spite of such differences, certain predicates have shown variation and change between the two patterns in recent times. The matrix predicate *submit*, in constructions involving subject control, is a case in point. It was shown that in the CEN both types were found and that *to* infinitives were much more common than *to -ing* complements. In the British English part, they were about five times more common, and in the American English part they were about twice as common.

It was also observed that the two variant types of sentential complement were found in one and the same text by one and the same author or in texts by one author that were only a few years apart. It is not easy to find a clear and consistent semantic distinction to explain the alternation. Further, if the notion of syntactic diaglossia is understood to refer to a difference in social register in the use of the variants, the present case is of interest because the alternants do not differ in this respect since they are found in the narration of the author in the texts. It is desirable to investigate the alternation in sentential complements selected by *submit* further when larger corpora of the 1880–1922 period become available, but it is suggested here on the basis of the variation found in the works of Humphry Ward and George Gissing that it may be necessary to allow for the availability, for individual speakers, of competing grammars that exhibit non-diaglossic variation during a time of grammatical change.

As for present-day English, this study shows that sentential complements involving subject control have become much more rare with *submit* than they were a century ago. However, such complements do still occur, and they exhibit a dramatic change. In both British and American English such complements are predominantly of the *to -ing* type.

The change of the argument structure properties of *submit*, it is argued in this study, can be seen as part of a larger pattern of changes affecting the system of English predicate complementation in recent centuries, which has been termed the Great Complement Shift. A number of aspects bearing on the Shift in the case of *submit* were pointed out, including the availability of *to* NP complements to

submit in the relevant sense. However, it was also argued, on the basis of an idea in earlier work, that the frequency of passive lower clauses may be a factor in the case of *submit*, promoting the emergence of the *to -ing* pattern.

This study also suggests that there is a need to examine the construction further in Canadian English. The segment of Canadian English that is included in the Bank of English Corpus is only a small corpus of 15.5 million words, and in that corpus, sentential complements of *submit* were surprisingly frequent and among such complements, *to* infinitives were also surprisingly frequent, but the numbers of tokens were low, and some of them came from one and the same text. The task is to examine the two types of sentential complements in a larger corpus, of perhaps at least 50 or a 100 million words.

Overall, the present study provides an illustration of the way that electronic corpora can shed light on the core grammar of English. It also suggests that there is a definite need in corpus studies today to compile larger and larger corpora of different regional varieties of the language and of different text types within such regional varieties.

6. Notes

* The author is grateful to Ian Gurney, of the English Department of the University of Tampere, for reading a preliminary version of this article and commenting on it. The author is also grateful to Elina Sellgren, likewise of the University of Tampere, and to Rena Torres-Cacoullos, of the University of New Mexico, for reading the article and commenting on it. The author is solely responsible for all shortcomings that remain.

1. For the relevance of extraction as a factor in general, see the pioneering study by Vosberg (2003). The present instance of extraction in (27a) is also of broader theoretical interest because it involves the extraction of an adjunct, a *for* phrase, out of a complement clause. This shows the need not to limit the consideration of extractions to the extractions of complements out of sentential complements, as was originally proposed by Vosberg (2003), and instead supports the idea of a broader view of extraction that also includes the extraction of adjuncts, as argued for by Rudanko (2006).

References

Allerton, D. (1988), "'Infinitivitis' in English," in: J. Klegraf and D. Nehls (eds.) *Essays on the English language and applied linguistics on the occasion of Gerhard Nickel's 60th birthday*. Heidelberg: Julius Groos, 11–23.

Chomsky, N. A. (1981), *Lectures on government and binding*. Dordrecht: Foris.

Chomsky, N. A. (1986), *Knowledge of language: its nature, origin, and use*. New York: Praeger.

Denison, D. (1998), "Syntax," in: S. Romaine (ed.) *The Cambridge history of the English language,* volume IV: 1776–1997. Cambridge: Cambridge University Press, 92–329.

Huddleston, R., and G. Pullum (2002), *The Cambridge Grammar of the English language*. Cambridge: Cambridge University Press.

Jespersen, O. (1961 [1940]), *A modern English grammar on historical principles*, part V: syntax (4th vol.). London and Copenhagen: George Allen and Unwin and Ejnar Munksgaard.

Kjellmer, G. (1980), "*Accustomed to swim*; *accustomed to swimming*: on verbal forms after TO," in: J. Allwood and M. Ljung (eds.) *ALVAR: a linguistically varied assortment of readings. Studies presented to Alvar Ellegård on the occasion of his 60th birthday.* Stockholm: Almqvist and Wiksell, 75–99.

Kroch, A. (2001), "Syntactic change," in: M. Baltin and C. Collins (eds.) *The handbook of contemporary syntactic theory.* Oxford: Blackwell, 699–729.

The Oxford English dictionary (1989). Oxford: Clarendon Press.

Pintzuk, S. (2003), "Variationist approaches to syntactic change," in: B. D. Joseph and R. D. Janda (eds.) *The handbook of historical linguistics.* Oxford: Blackwell, 509–528.

Poutsma, H. (n.d.) *Dictionary of constructions of verbs, adjectives, and nouns.* Unpublished manuscript. Copyright: Oxford University Press.

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik (1985), *A comprehensive grammar of the English language*. London: Longman.

Roberts, I. (2007), *Diachronic syntax*. Oxford: Oxford University Press.

Rohdenburg, G. (2006), "The role of functional constraints in the evolution of the English complementation system," in C. Dalton-Puffer et al. (eds.) *Syntax, style and grammatical norms: English from 1500–2000*. Bern: Peter Lang, 143–166.

Ross, J. (1973), "Nouniness," in: O. Fujimura (ed.) *Three dimensions of linguistic theory*. Tokyo: TEC Company, 137–257.

Rudanko, J. (1998a), *Change and continuity in the English language*. Lanham, MD: University Press of America.

Rudanko, J. (1998b), "*To* infinitive and *to -ing* complements: a look at some matrix verbs in Late Modern English and later," *English studies*, 79: 336–349.

Rudanko, J. (2006), "Watching English grammar change: a case study on complement selection in British and American English," *English language and linguistics*, 10(1): 31–48.

Vosberg, U. (2003), "The role of extractions and *horror aequi* in the evolution of *-ing* complements in Modern English," in: G. Rohdenburg and B. Mondorf (eds.) *Determinants of grammatical change in English*. Berlin: Mouton de Gruyter, 305–327.

Wasow, T., and T. Roeper (1972), "On the subject of gerunds," *Foundations of language*, 8: 44–61.

Journalistic corpus similarity over time

*Cristina Mota**

New York University and Instituto Superior Técnico and L2F/INESC-ID

Abstract

We used the method proposed in Kilgarriff (2001) to assess corpus similarity over a short period of time both within topic and cross topic. The corpus samples were drawn from a Portuguese journalistic corpus. The corpus spans eight years (from 1991 to 1998) and comprises article extracts marked with the year segment, half-year segment, and newspaper section of publication. We analyzed the corpus, taking as reference each text in the time interval and comparing it with all texts published in different periods. We observed that (i) the similarity between two texts within the same topic generally decreases as the time gap between them increases, being more significant for some topics, and (ii) in some cases the texts on one topic over time become as different as two texts from different topics. Since the ultimate goal of our work is to understand how the changes in corpus similarity affect the performance of a named entity tagger, we also measured similarity based on frequency lists containing only capitalized words and containing only lowercase words. The former similarity aims at comparing the corpora from the viewpoint of the named entities content, whereas the latter one approximately compares the surrounding contexts of the named entities. The results show that the similarity values based on these lists also generally decrease over time, even though the decreasing profiles are topic-dependent.

1. Introduction

Language is dynamic; it changes in many different ways, accompanying political, social, and cultural trends; adapting to the evolution of science and technology; and even suffering the influence of other languages and foreign speakers. However, finding evidence that some level of language has changed is not an easy task for a linguist and requires large amounts of corpora spanning several years. In order to detect lexico-grammar and grammar changes, for example, as Renouf (2002) points out, one decade of corpora is insufficient to study such phenomenon.

From the point of view of natural language processing (NLP) systems, which model only a snapshot of language represented in the form of written or spoken corpora, language changes very rapidly. Systems are presented frequently with linguistic objects they have never seen before (i.e., objects that do not fit the language model built from previous data). This factor is of particular relevance for systems that process texts with streamlike characteristics, such as news, scientific papers, or even blogs, where information about a certain topic appears suddenly, is very common during a period of time, and then tends to disappear or come back again some time later.

Both for language and systems, this perspective is fundamentally diachronic. At a certain point in time, texts may present certain characteristics that may not exist in either earlier or later texts. Our main objective is to demonstrate that these changes altogether represent over time a decreasing trend in shared properties between texts that affects the performance of NLP systems in an identical way.

We opted for focusing our study on named entity recognition for journalistic texts.[1] This task could be particularly sensitive to the temporal distance between texts used to train and test the system, because proper names occurring in the news strongly depend on the geopolitical, sociocultural, and technological situation they are reporting. Intuitively, the more two texts are temporally distant, the less likely it is that a certain name (e.g., the name of a country's president or company's CEO) is common in both texts. In that case, in order to be able to recognize names they are observing for the first time, systems have to rely more on external evidence, a factor that could also be subject to variations due to changes in time period.[2]

Although our main objective is related to how the performance of named entity recognition varies over time, we first addressed the question of whether texts over time do indeed share fewer properties, thus becoming less similar, which is the main focus of the current paper. This issue was investigated by applying the method proposed by Kilgarriff (2001), but instead of comparing language varieties, we compared samples drawn from different time periods. The study was conducted in Portuguese using samples drawn from a journalistic corpus with 180 million words spanning eight years of news from 1991 to 1998.

In this paper, we provide evidence that in a short period of eight years, samples drawn from 16 consecutive six-month periods become less similar as we increase the time gap between the samples being compared. The analysis was done using not only the most frequent words but also the most frequent capitalized words and the most frequent lowercase words separately.

We begin, in Section 2, by briefly presenting the approach we adopted to compare the corpus over time, framing it in the context of related work. In Section 3, we characterize the corpus used in our study from the point of view of the vocabulary growth curves, both by topic and time period. Then, in Section 4, we motivate and describe in detail different analyses and corresponding results. We conclude by summarizing and discussing the main results.

2. Approach to corpus similarity

Our very general hypothesis is that the performance of named entity recognition decreases as we increase the time gap between the texts used to train and test the system, as a result of those texts becoming less similar over time.

In order to test this, we need a comparison method that relies on nothing but words, which can be more easily recognized than other linguistic units, such as multiword expressions, named entities, or syntactic constituents. There are two

main reasons for imposing this requirement: (i) texts manually annotated with linguistic information (such as part-of-speech tags, parse trees, named entities, etc.) are expensive to obtain; (ii) using an NLP system to perform such tasks most likely requires manually annotated information to train the system (which leads us back to reason [i]), or, in other words, even if the need for labeled information is minimal (when using semi-supervised or unsupervised methods), the resulting annotation could skew our corpus analysis over time. This bias would be a consequence of what we want to ultimately show: how the performance of NLP systems, in particular of named entity taggers, is affected by increasing the time gap between texts. This means that either the system produces erroneous results or is not capturing all the information it should.

Furthermore, we want a method to be useful for comparing texts before training a system (with unlabeled data) for a specific NLP task. The objective is to select the training texts that are the most similar to the texts used to test the system, thus enhancing the performance of the system on those texts.

The method proposed by Kilgarriff (2001) to compare similarity between language varieties was a perfect fit to our purposes. On the one hand, as Kilgarriff clearly stated, "Reliable statistics depend on features that are reliably countable and, foremost amongst these, in language corpora, are words" (233), and on the other hand, as the author also points out, it is a practical question to understand how costly it is to port an NLP system from one domain to another, given corresponding corpora.

Kilgarriff measures corpus similarity based on the distance between word frequency lists. The author argues that the similarity results can only be properly interpreted with respect to homogeneity, defining homogeneity/heterogeneity as the within-corpus distance (i.e., distance between two halves of the same corpus), and similarity/dissimilarity as the distance between two different corpora (i.e., distance between two halves of two different corpora). In other words, looking solely to one of the values (within-corpus distances or distance between two corpora) is useless; the distances need to be compared to each other to be able to conclude anything meaningful about the similarity.

In fact, Baayen (2001: 34) advocates an equivalent position:

> In addition, to gauge the importance of a difference in the value of a text characteristic for two or more texts, one should weigh the intertextual differences with respect to the intratextual variability of the text characteristic. It is only when the intertextual differences are larger than the intratextual differences that one may have some confidence that the differences are reliable.

We should stress that although the subject of corpus homogeneity and similarity is understudied, there is already a significant body of work in this area. The reader should refer to Gries (2006) for a very complete description and discussion of related work and also for a proposal for assessing variability within and between corpora based on resampling methods and exploratory data analysis.

Furthermore, the issue of whether variability represents a relevant fact is also important, and the use of null hypothesis testing (by means of the χ^2 statistics, for instance) to analyze variability has been often criticized due to the non-random nature of language, namely by Kilgarriff (2005). In a follow-up article, Gries (2005) further investigates this problem using effect sizes.

2.1 Kilgarriff's corpus homogeneity-similarity methodology

Given two corpora, *A* and *B*, with the same number of words, Kilgarriff measures the similarity between *A* and *B*, *Similarity*(*A*,*B*), by applying the following algorithm:

1. Split corpora *A* and *B* into *k* slices each;
2. repeat *m* times:
   a. randomly allocate *k*/2 slices from *A* to A_i and *k*/2 slices from *B* to B_i
   b. construct word frequency lists for A_i and B_i
   c. compute distance between word frequency lists of A_i and B_i for the *n* most frequent words of the joint corpus A_i+B_i
3. output mean and standard deviation of the distances obtained over all iterations i=1…m

After measuring *Similarity*(*A*,*B*), one must also compute *Similarity*(*A*,*A*) and *Similarity*(*B*,*B*), which give the within-corpus similarity (i.e., corpus homogeneity) of corpora *A* and *B*, respectively.

In his study, Kilgarriff compared different distance metrics by applying the algorithm to Known-Similarity Corpora (KSC). As χ^2 by degrees of freedom (CBDF) (i.e., χ^2 normalized by the degrees of freedom) performed best on the KSC, we used this metric in our study.

2.2 Corpus similarity over time

In our study, instead of comparing language varieties, we compared texts from different time periods. This means that given a time interval *T* segmented into *N* time periods of equal size, we measured *Similarity*(t_i,t_j) where t_i and t_j are texts drawn from time periods i=1...N and j=1...N.

The interpretation of the similarity results is done by using as reference the values for *Homogeneity*(t_i), i.e., the within-similarity for texts drawn from the same time period. These values, as we will see, correspond to the minima of all values.

We also used the same method to compare texts drawn from each of the five topics available in our corpus, and beyond basing the comparison on word frequency lists, we replicated the analysis using separate lists of lowercase and capitalized words.

3. Corpus characterization

In our study we used subsets of the raw version 1.7 of CETEMPúblico (Rocha and Santos, 2000). This corpus is the largest Portuguese public journalism corpus, covering eight years of news, from 1991 to 1998, with a total of 180 million words. CETEMPúblico comprises articles from about 2,600 daily editions of the Portuguese newspaper *Público* that were fragmented into extracts (about two paragraphs each) and randomly shuffled within each half-year segment by the corpus builders. This means that the order of the extracts in the corpus may not correspond to the chronological order of the extracts and that two adjacent extracts are unlikely to belong to the same original article.

It should be noted that one of the base data sets used to build the corpus covered articles from 1991 to 1995, and it comprised files containing a complete daily edition of the newspaper. The 1991 articles were not classified, and the articles from 1992–1995 were classified according to the newspaper layout (for instance, *Last Minute*); the other data set covered articles from 1996–1998, and it comprised files containing a single article classified according to a new layout. These differences in formatting (which forced the corpus builders to heuristically segment the complete editions into articles) and in classification may explain the higher vocabulary diversity of texts belonging to the first period than of texts from the second period.

The corpus classification (which differs from the original newspaper layout) consists of nine subjects, henceforth designated topics, but we analyzed only Culture, Sports, Economy, Politics, and Society, because these are the only topics that are covered by all time periods. Besides restricting our analysis to five topics, we opted for processing just sentences (i.e., titles, authors, and list elements). We kept the original tokenization of the corpus, where each token is on a separate line.

In order to characterize the corpus in more detail by topic and time period, prior to conducting our analysis over time, we adopted the framework proposed by Baayen (2001) and the zipfR module by Evert and Baroni (2007). We were particularly interested in the vocabulary growth curves, which clearly show the rate at which the vocabulary increases as we increase the size of the corpus. Furthermore, based on those curves, we also estimated the lexical productivity, which is the rate of unique words,(i.e., hapax legomena) relative to the size of the sample, since samples of equal size to identify which topics and time periods had higher rates of unique words.

The complete corpus was sampled by increasing the sample size by 1,000 words at each step i in two different ways: (i) *by topic*, where each topic was sampled chronologically by time period and the original order of the extracts within the corpus was kept within the same time period, and (ii) b*y topic and time period*, where the extracts belonging to the same topic and time period were sampled, preserving the original order within the corpus. At each sampling step i, we measured the number of words, N_i; the vocabulary size of the sample, $V(N_i)$; and the number of hapax legomena in the sample, $V(1,N_i)$. Moreover, we also

measured the same metrics based on lowercase words (i.e., words with no uppercase letters), capitalized words (i.e., words not occurring at the beginning of sentence that have mixed case and start with an uppercase letter), and uppercase words (i.e., words not occurring at the beginning of sentence that have only uppercase letters). In the final step, we created the term frequency list for each kind of word.

3.1 Topic perspective

Figure 1 shows the vocabulary growth curves for the four word types; each curve in one graph corresponds to a different topic. Given two samples with the same size N_i,, the larger the vocabulary size of a sample $V(N_i)$, the more diverse the words of that sample, i.e., the smaller is the mean word frequency $N_i/V(N_i)$.

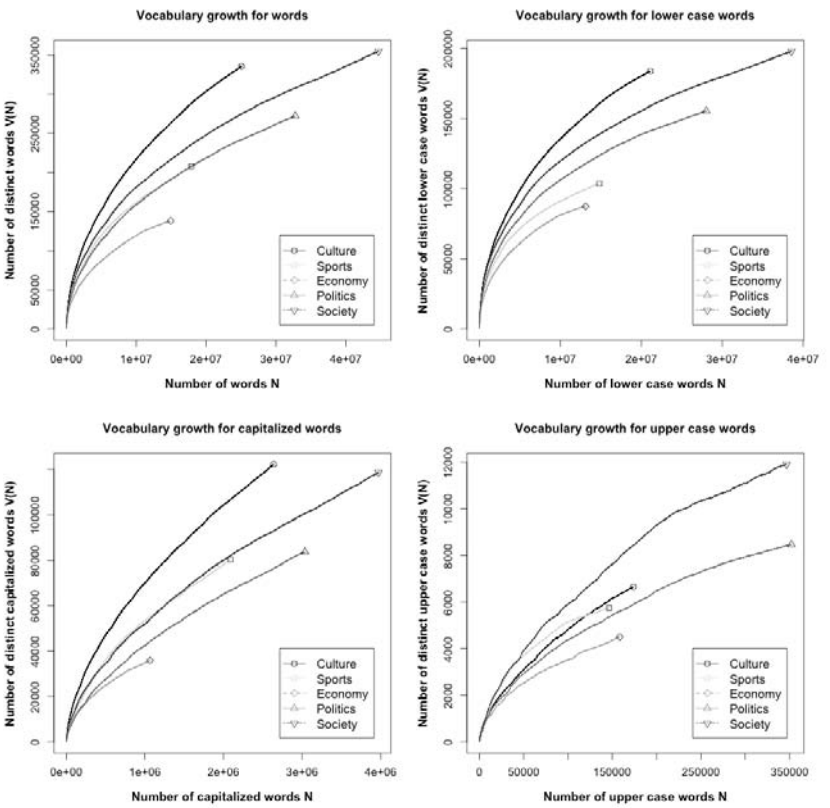


Figure 1. Empirical vocabulary growth curves by topic and word type.

As can be seen, Economy is the topic with the least diversified vocabulary for all word types, whereas Culture is the topic with the most diversified vocabulary, except for uppercase words. For uppercase words, Society is the most

diversified topic. In fact, the lexical productivity values in Table 1 confirm this observation and show that capitalized words are clearly the most productive word type (on average, about seven times more productive than lowercase words).

Table 1. Lexical productivity $E[V(1,N_i=3{,}585\text{K})]/N_i$ by topic and word type.[3]

	Word type				
	Culture	Sports	Economy	Politics	Society
Words	0.00389	0.00280	0.00160	0.00284	0.00338
Lowercase	0.00237	0.00127	0.00108	0.00169	0.00201
Capitalized	0.01570	0.01336	0.00526	0.01107	0.01412
Uppercase	0.01016	0.00964	0.00428	0.01094	0.01356

The vocabulary growth curves for lowercase words are the only curves that distinguish the topics in terms of word diversity; for words, the vocabulary size of Sports and Politics increase at a similar rate (the curves are overlapped), while for capitalized words, the vocabulary size of Sports increases at a similar rate as the vocabulary size of Society.

Furthermore, it is also apparent that the curves for uppercase words, especially for Society, seem to change shape after a certain sampling step. Since the time periods of the topics are ordered chronologically, this change could correspond to the start of the period 1996–1998. This issue will be clarified below.

3.2 Topic and time period perspective

The analysis of the vocabulary growth curves by time period within each topic for words and lowercase words shows that the topics are lexically more varied in the period 1991–1995 than in the period 1996–1998 (with the exception of the first half of 1991, which has a variety closer to the second period). The distinction is particularly visible in Culture, Economy, and Politics (Figure 2 shows the graph for Politics), where a clear gap between the curves in the periods 1991–1995 and 1996–1998 can be observed. In Sports and Society, such a gap is not visible, but the curves for the period 1991–1995 are above the curves for the period 1996–1998.

4. Analysis over time

The corpus used in our study is organized into topics and time periods that span six months each. Within each six-month period, the extracts are randomly shuffled, and besides the year, they do not contain finer timestamps, such as day or month of publication. Therefore, as previously mentioned, the minimum time unit of analysis is six months. We analyzed the corpus over time by comparing extracts from each time period in the time interval with extracts drawn in turn from all time periods in the same interval of eight years. This corresponds to 256 comparisons (16 half-year segments x 16 half-year segments) per topic.

The samples used in our analysis correspond to the first 4,140 extracts within each half-year segment and topic;[4] each sample is limited to 370,000 words.[5] We conducted the similarity analysis over time within each topic and also across topics. Initially, we compared word frequency lists and then replicated the analysis using frequency lists of lowercase and capitalized words separately.

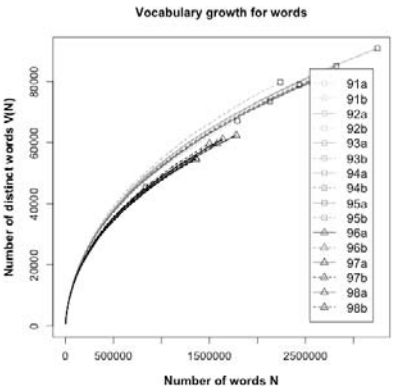


Figure 2. Empirical vocabulary growth curve for Politics by time period.

In this section, we use the following terminology and notation, where t, t_1, and $t_2 \in$ {Culture, Sports, Economy, Politics, Society} and i,j=*91a,…, 98b*:

- $T_{t,i}$ is a set of extracts drawn from topic t and time period i.
- *Homogeneity*($T_{t,i}$) is the homogeneity for topic t in time period i.
- *Similarity*($T_{t,i}$,$T_{t,j}$) is the within-topic similarity over time for topic t.
- *Similarity*($T_{t1,i}$,$T_{t2,j}$) is the cross-topic similarity over time between topics t_1 and t_2.

In all cases, the number of slices k was established at 10, and the number of most frequent words 2,000. In the analysis using lowercase words and capitalized words, we used 1,600 and 160 words, respectively. These values were arbitrarily chosen to be 80% and 8% of n.

4.1 Within-topic homogeneity

According to Kilgarriff (2001), as discussed in Section 2, the similarity between two corpora can only be interpreted relative to the within-corpus distance of each corpus (i.e., relative to their homogeneity). Hence, prior to beginning the analysis of the similarity over time and cross topic, we show in Figure 3, for each topic t, the corresponding boxplots of the within-corpus distance for all half-year segments in each topic. Each boxplot summarizes the values obtained for *Homogeneity*($T_{t,i}$), which did not present any visible trend within each topic.

As can be seen, the within-corpus distance values vary close to 1, with Politics having the lowest values, which means it is the most homogenous topic, while Culture is the least homogeneous of the five topics; it is also obvious that Sports spreads more widely. This result contrasts with the vocabulary growth curves of Figure 1 and the values for the lexical productivity in Table 1 that showed that the topic with the lowest lexical productivity was Economy. In any case, the topic with the highest lexical productivity, Culture, is the least homogeneous. This means that the level of homogeneity cannot be directly assessed from the values of the lexical productivity.

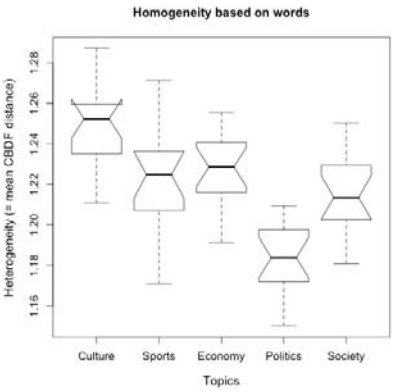


Figure 3. Boxplots of the within-corpus distance obtained for each topic.

4.2 Within-topic similarity over time

We have just seen that the within-corpus distance values vary around 1, which will serve as a reference point for the current analysis. The question we will address now is whether the similarity decreases (i.e., the distance values increase) compared to this reference as we increase the time gap between the texts being compared. The graphs in Figure 4 show the similarity for each topic over time $Similarity(T_{t,i},T_{t,j})$, where the points in time gap 0 correspond to the homogeneity $Homogeneity(T_{t,i})$ shown in the previous section; the boxplot in Figure 4 (top left graph) summarizes the results.

The first visible result is that comparing texts with different time periods yields higher distance values, as shown by the boxplots in Figure 4; the minima are also around 1, which correspond to the homogeneity values, but the maxima vary between 2.77, in the case of Society, and 6.81, for Economy. Again, it is interesting to observe that there is not an association between the similarity over time for one given topic and the homogeneity level of that topic, (i.e., the fact that the texts drawn from different time periods within the same topic are more homogenous does not necessarily mean that their similarity is higher over time) (see Figure 3 with top left graph in Figure 4).

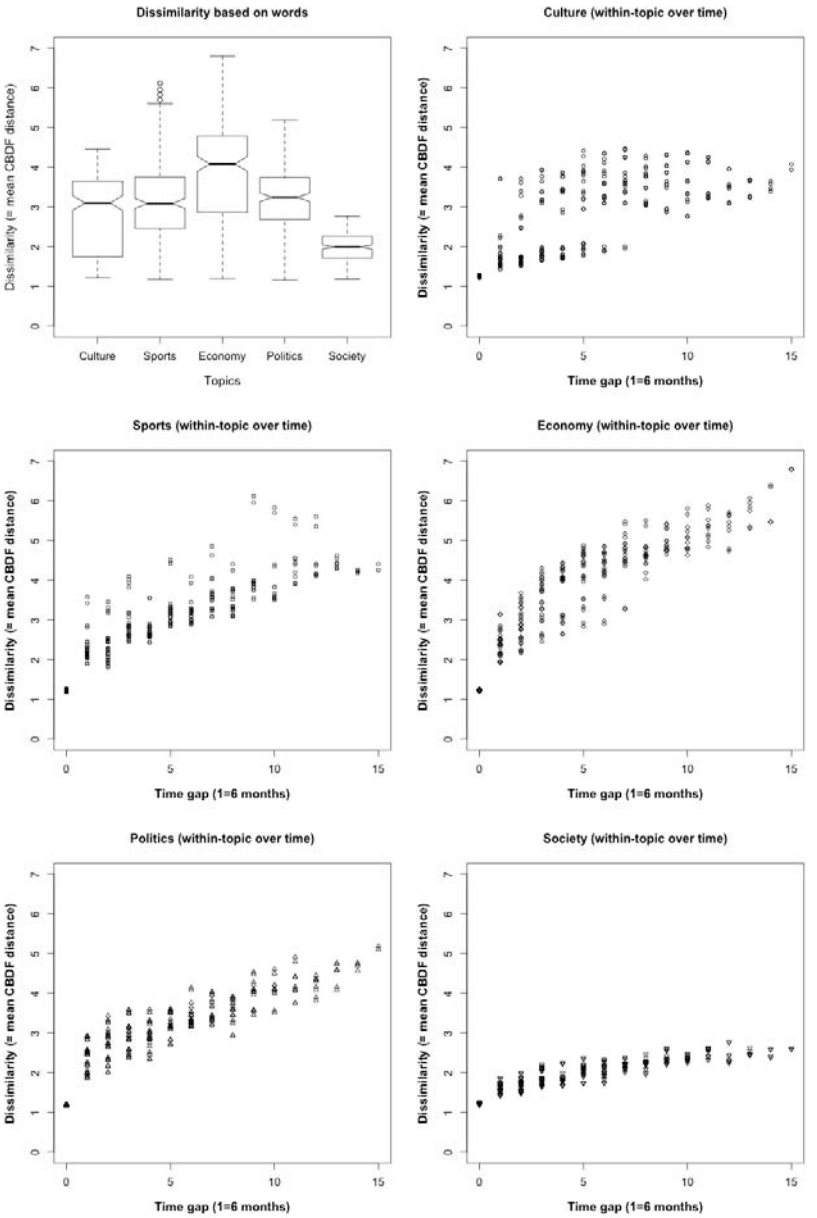


Figure 4. Within-topic similarity over time for each topic.

For instance, Economy and Society have the same degree of homogeneity, but Society presents significantly smaller distance values between texts with different time periods than Economy.

The second important result is that all topics present higher distance values as the time gap between texts becomes larger (i.e., there is a decreasing similarity trend over time, as seen in all graphs of Figure 4). For example, in Politics (see bottom left graph), when the time gap between texts is 0, the corpus distance values vary between 1.150 and 1.209, but increasing the distance by one year (time gap 2) makes the corpus distance values range from 2.003 to 3.435; and when the time distance is six years (time gap 12), the dissimilarity values spread between 3.817 and 4.454.

We notice, however, that the similarity does not strictly decrease; sometimes the minimum and/or the maximum corpus distances slightly decrease when the time gap is increased (see values for time gaps 4 and 5 in *Politics* or time gaps 11 and 12 in *Economy*). This behavior could be due to the fact that for some time gaps we have less points, or it could be a reflection of certain cyclic events in the news (elections, Olympics, playoffs, terrorist attacks, scandals, political crises, etc.) or the appearance of new transient events that may cause greater increases in the distances than expected and then, after disappearing, cause the inverse effect.

Finally, it is also worth pointing out that the similarity curves do not show a tendency to flatten, a result that suggests that if the time span were larger than eight years, the distance between texts would possibly continue to increase.

4.3 Cross-topic similarity over time

In order to fully interpret the similarity results over time, we also investigated how texts drawn from one topic compare to texts drawn from other topics. Hence, for each topic we analyzed the $Similarity(T_{t1,i}, T_{t2,j})$. In this case, we wanted to observe whether the similarity changes over time when comparing texts from different topics and also to have reference values against which to compare the within-topic distance values over time. Although some topics may share news in the same time period, we are not expecting a decreasing trend, and we are expecting the cross-topic distance values to be much higher than within-topic distance values over time. In this section, we illustrate the results of comparing Politics with the other four topics.

As can be observed in the boxplot of Figure 5 (left graph), the cross-topic distances between Politics and each of the other topics are higher than the within-topic distances over time, which are also plotted in the same figure, showing that Politics is more similar to Society (distance values between 7.771 and 12.100) and Culture (distance values between 10.66 and 16.17) than to Economy (distance values between 13.90 and 20.73) and especially Sports (distance values between 15.67 and 22.72). These results confirm our intuition that Politics should differ more markedly from Sports than from Society, a topic with which it is more likely to share some news.

In terms of the shape of the curves (see Figure 5, right graph), no particular decreasing trend in similarity is visible, and the values of the distance for each time gap spread more widely than when comparing texts drawn from

Politics. In fact, contrary to what happens with the within-topic similarity for each topic, the shapes of the cross topic are quite different from one to the other.

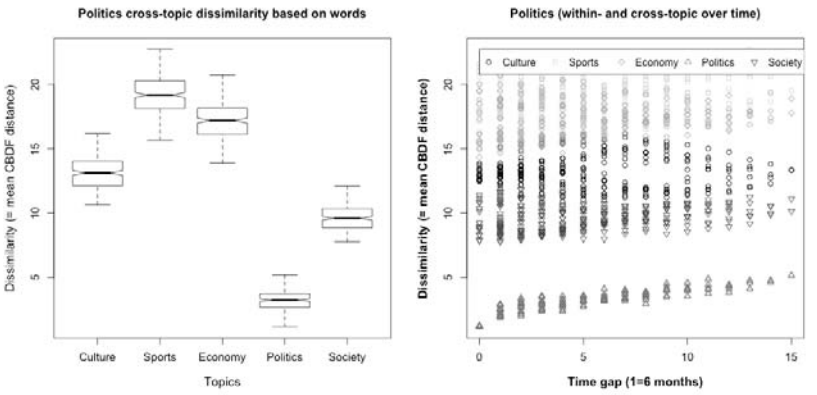


Figure 5. Within-topic and cross-topic similarity for Politics.

Finally, it is interesting (and unexpected) to observe that the within-topic similarity values over time approach the cross-topic similarity values (i.e., the values obtained by computing the similarity over time within one topic approximate the values obtained when comparing texts from different topics). For instance, the minimum value when comparing texts from Politics with Society is 7.771, whereas the maximum value obtained when comparing texts drawn from Economy at the maximum time gap is 6.806. This result implies that if the time interval was two or three years longer, texts drawn from Economy could show a degree of dissimilarity over time comparable to the degree of dissimilarity of texts belonging to different topics.

4.4 Within-topic similarity using lowercase and capitalized words

Given that the ultimate purpose of our work is to compare the performance of a named entity tagger over time and relate it with the corpus similarity over the same period, we replicated the within-topic analysis measuring corpus similarity based on: (i) capitalized words (in this case uppercase words also count as capitalized words) and (ii) words with lowercase only. The first similarity can be seen as a rough way of comparing the content of named entities (which mainly comprise capitalized words), whereas the second value aims at quantifying the similarity between the surrounding contexts of the named entities (which amainly comprise lowercase words). Instead of using the 2,000 most frequent words, we used the 160 most frequent capitalized words and 1,600 most frequent lowercase words in each of the cases, respectively.

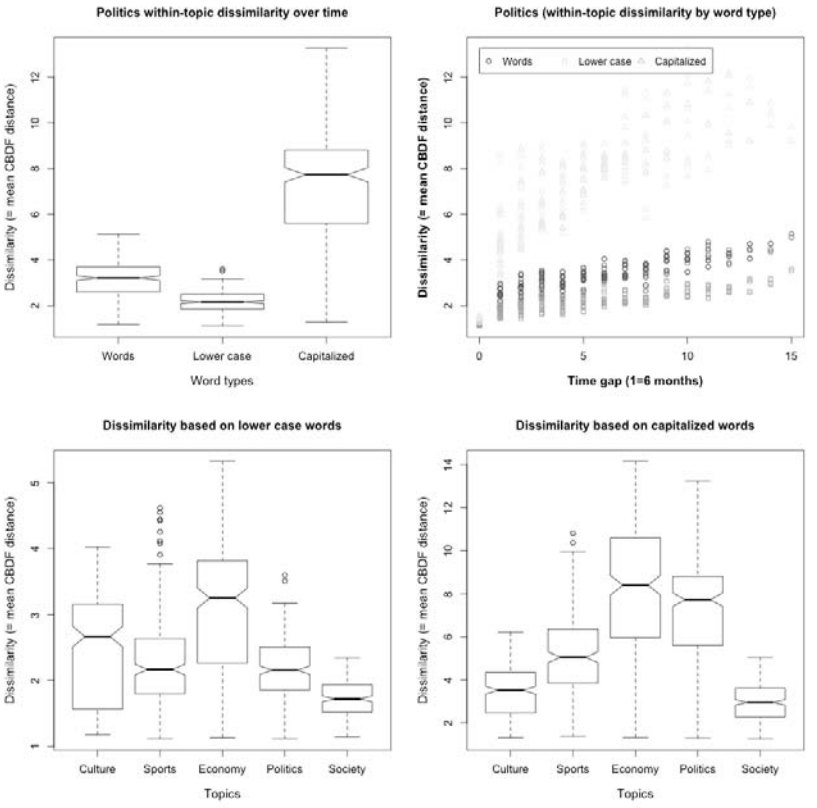


Figure 6. Within-topic similarity over time based on words, lowercase words, and capitalized words.

In the boxplot of Figure 6 (top left graph), we can see that measuring similarity with capitalized words within Politics yields higher corpus distance values (between 1.279 and 13.250) than with all words (between 1.169 and 5.128), which is expected, since it is natural that the largest contribution to the decrease in similarity over time based on all words comes from capitalized words. In fact, this explains the slightly lower distance values for the similarity based on lowercase words, which varies between 1.111 and 3.599.

In any case, regardless of the word type used to build the frequency lists, the similarity between texts decreases as the time gap between the texts increases (see Figure 6, top right graph). We show the corresponding boxplots for lowercase and capitalized words in the bottom graphs of Figure 6.

The similarity profiles depend on the topic, as can be seen by comparing the boxplots in Figure 6 with Figure 4 that summarizes the results using all word types. For instance, for all word types, Politics has distance values slightly higher than Culture, but with lowercase words it has slightly lower values; on the other

hand, using only capitalized words, we obtain much higher distance values over time for Politics than for Culture. This means that the decrease in similarity over time for some topics is more significantly affected by variations in the capitalized words than others.

Referring back to Section 3, this result shows that even though the lexical productivity of capitalized words for Culture is higher than for Politics, the change in the frequency of those words is higher in Politics.

5. Discussion

Our main goal was to verify that corpus similarity decreases over time, a factor that could impact the performance of named entity recognition. We observed that corpus similarity based on frequency lists of words, capitalized words, or lowercase words, decreases as we increase the time gap between the texts being compared. This means that as we increase the time gap between a reference text and other texts in the time interval, the frequency of words in those texts increases and/or decreases relative to the frequency of the same words in the reference corpus. The comparison was done for Culture, Sports, Economy, Politics, and Society in a Portuguese journalism corpus. Although all topics have close homogeneity values around 1, increasing the time gap results in higher values of dissimilarity depending on the topic and also on the type of words. The decreasing curves did not flatten within the period of eight years we studied.

These results can be interpreted as confirmations of null hypotheses, as it is certainly not random that news articles six months apart have more similar word frequency profiles than texts one year apart. However, from our point of view, it is not obvious that, for instance, news from 1998 still has more in common with news from 1992 than from 1991. In fact, this raises the question of how much one would need to increase the time gap between texts to stop observing such decrease in similarity, assuming that corpus similarity eventually stabilizes over time.

We also observed that texts within some topics over time were becoming as dissimilar as texts drawn from different topics. This finding also deserves further investigation, especially in the context of training a named entity tagger. Most researchers would find it inappropriate to train a named entity tagger aiming at texts dedicated to a particular subject—for instance, a named entity tagger for Economy, with texts dedicated to a completely different subject, such as Culture. If texts over time become as dissimilar as texts from different topics, then one should also avoid training a named entity tagger with texts that are temporally too distant from the texts the tagger is going to process.

Concerning this issue, Mota and Grishman (2008) assessed the impact of these decreasing trends on the performance of named entity recognition. They showed that for the Politics section of the same corpus, there is a negative correlation between corpus similarity and the performance of named entity recognition: the performance decays as a result of increasing the time gap

between training and test data, which is related to a decrease in similarity between those texts.

In that case, and if the performance profiles are directly related to the similarity profiles, one should expect that the performance of a name tagger for Economy is the most affected by increasing the time gap between the training and test texts, since those are the texts that became the least similar over time.

Regarding the comparison based on other word types, the comparison based on capitalized words shows a more significant decrease in similarity over time than based on lowercase words. This is a consequence of a larger variation in the frequency of capitalized words over time than in the variation of lowercase words. Despite not using the same number of most frequent words when comparing frequency lists containing different word types, the comparison between the decreasing curves is meaningful, because the level of homogeneity based on those word types is close to the same value, so they all start from around the same minima at time gap 0.

We made an effort to create the most appropriate experimental conditions from which we could conclude that increasing the time gap was the main reason for observing a decrease in corpus similarity. In particular, we analyzed a corpus from one single newspaper and used samples of equal size for each time period and topic. However, two factors may have introduced some noise to the analysis. On the one hand, the corpus was built from two sets of data with different properties, especially regarding the classification of the articles, and on the other hand, texts within the same six-month period are not chronologically ordered because they correspond to fragments of the original articles that were randomly shuffled. The former factor may have been responsible for observing different lexical properties within three periods—1991, 1992–1995, and 1996–1998; the latter characteristic of the corpus prevented us from having more fine-grained time periods and also makes it almost impossible to know whether the texts of one sample are days, weeks, or months apart. Both factors could be pointed to as a source of some unexpected variations. Of course those variations could be simply a result of sudden shifts in the news or a combination of the two.

The fact that we are using a corpus built with articles from a single newspaper creates better conditions, but still has a major drawback. It raises the question of whether the same kind of trends would be observable if one had used a more diverse corpus. We believe that as long as the sources were consistent for each time period, such as the proceedings of a series of conferences, similar results could have been achieved. In any case, the problem we studied is mainly relevant to processing streamlike texts, such as broadcast news, newswire, newspapers, proceedings, journals, or web logs.

Using the same amount of data (in terms of words) is also a limitation, because in order to do that, we had to truncate the data given the size of the smallest set, and the corpus has very different sizes per time period and topic. So, instead of analyzing 180 million words, we ended up using only about 20% of the corpus in our study.

6. Notes

* This research was partly funded by Fundação para a Ciência e a Tecnologia through a doctoral scholarship (ref: SFRH/BD/3237/2000). We are very grateful to Ralph Grisham for reviewing this paper several times, to Adam Kilgarriff for his prompt support when we first attempted to implement his method, and also to the paper reviewers for their valuable comments and suggestions.

1. According to the task definition proposed in the 6th *Message Understanding Conference* (MUC-6), named entity recogniton is the idenfication and classification of proper names of people, organizations, locations, and other related objects, such as temporal and numerical expressions (see Grishman and Sundheim [1995]).
2. We adopt the term external evidence from McDonald (1996), which refers to the context surrounding the name or other clues not relative to the name itself that allow to classify a name. An example of external evidence is the verb of which the name is the subject. For instance, in *John swims every weekend*, *swims* gives evidence that *John* is the name of a person.
3. The expected vocabulary size for the hapax legomena was estimated, for a sample size of 3,585,000 words (which is the size of the largest topic) using the finite Zipf-Mandelbrot model as implemented in the ZipfR module with the default parameters.
4. This number of extracts (4,140) is slightly smaller than the minimum number of extracts given all complete sets of extracts in one half-year period and topic.
5. A sample is limited to 370,000 words because it is comprised of 10 slices containing each 37,000 words; 37,000 is the size of the smallest of the 800 slices (which correspond to 10 slices x 5 topics x 16 time periods) used in all comparisons.

References

Baayen, R. H. (2001), *Word frequency distributions*. Dordrecht: Springer.

Evert, S., and M. Baroni (2007), "zipfR: word frequency distributions," in: A. Zaenen and A. van den Bosch (ed.) *Proceedings of the 45th annual meeting of the ACL: companion volume proceedings of the demo and poster sessions*. ACL, 29–32. Online at: <http://www.cogsci.uni-osnabrueck.de/~severt/PUB/EvertBaroni2007.pdf>.

Gries, St. Th. (2005), "Null-hypothesis significance testing of word frequencies: A follow-up on Kilgarriff," *Corpus linguistics and linguistic theory,* 1(2): 277–94.

Gries, St. Th. (2006), "Exploring variability within and between corpora: Some methodological considerations," *Corpora,* 1(2): 109–151.

Grishman, R., and B. Sundheim (1995), "Design of the MUC-6 evaluation," in: *Proceedings of the 6th message understanding conference*. San Mateo, CA: Morgan Kaufmann. Online at: <http://www.aclweb.org/anthology/M/M95/M95-1001.pdf>.

Kilgarriff, A. (2001), "Comparing corpora," *International journal of corpus linguistics,* 1(6), 1–37.

Kilgarriff, A. (2005), "Language is never, ever, ever random," *Corpus linguistics and linguistic theory,* 1(2): 263–276.

McDonald, D. D. (1996), "Internal and external evidence in the identification and semantic categorization of proper names," in: B. Boguraev and J. Pustejovsky (eds.) *Corpus processing for lexical acquisition*. Cambridge: MIT Press, 21–39.

Mota, C., and R. Grishman (2008), "Is this NE tagger getting old?" in: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, and D. Tapias (eds.) *Proceedings of the 6th international language resources and evaluation conference*. ELRA, 1196-1202. Online at: <http://www.lrec-conf.org/proceedings/lrec2008/pdf/303_paper.pdf>.

Renouf, A. (2002), "The time dimension in modern corpus linguistics," in: B. Kettemann and G. Marko (eds.) *Teaching and learning by doing corpus analysis*. Amsterdam: Rodopi, 27–41.

Santos, D., and P. Rocha (2001), "Evaluating CETEMPúblico, a free resource for Portuguese," in: *Proceedings of the 39th annual meeting of the ACL*. ACL, 454–457. Online at: <http://www.aclweb.org/anthology/P/P01/P01-1058.pdf>.

"Ah lovely stuff, eh?"—invariant tag meanings and usage across three varieties of English

*Georgie Columbus**

University of Alberta, Canada

Abstract

Invariant tags, such as huh *and* innit*, are discourse markers that often occur at the end of an utterance to provide attitudinal and/or evidential information above that of the proposition. Many previous studies examined the meaning or usage of these tags in single varieties or dialects of English. Few of these studies, however, have examined variation in invariant tag use. Some studies have investigated sociolinguistic divisions within a dialect, but none have compared usage between varieties. Furthermore, differences in research methodology and aims prevent comparison of the prior results. This study investigates the meaning/functions of four invariant tags—*eh*,* yeah*,* no*, and* na*—in New Zealand, Indian, and British English. The four most frequent meanings are described in detail. The results show differences in the meanings available as well as in their usage frequencies across both items and varieties. This suggests that varietal differences at the level above propositional understanding could cause problems for intercultural and global communication. This has implications for pedagogy and materials for English for Speakers of Other Languages (ESOL) and English for Specific/Business Purposes, in that global communication in English requires an awareness of these subtle differences at the varietal level.*

1. Introduction

Discourse markers, such as *you know*, *like*, and *well*, have long held the attention of linguists, and the advent of corpus-based data mining has allowed detailed description of trends and usage. One particular type of discourse marker is the Invariant Tag (InT). These tags are similar to canonical question tags (i.e., *isn't it?*, *do they?*), in that they share some of the verification functions of canonical tags, but these InTs have broader applications. Previous research in discourse particles such as these, however, has often been limited to one variety or dialect (e.g., Stubbe and Holmes, 1995, in New Zealand English; Schiffrin, 1987, in U.S. English) or to one marker within a dialect/variety (e.g., Norrick's 1995 study of U.S. *huh,* which he spells *hunh*; Tagliamonte and D'Arcy, 2004, on Toronto *like*). If and to what extent these definitions and descriptions are applicable to other varieties or dialects of the same language remains unclear.

1.1 Background on tags

Question tags in their canonical form have been much studied for their curious semantic and syntactic properties, such as polarity and agreement. Holmes's

(1982) study of New Zealand English (NZE) question tags is a thorough classification of the functions available to NZE speakers. Functional categorizations such as these have resulted in many clear descriptions in English grammars, particularly for ESOL purposes. InTs, on the other hand, have no polarity or agreement quirks, and may share their orthographic form with words that act as regular adjectives, interjections, adverbs, or even other discourse marking items.[2] Studies on InTs are numerous, and most delve into sociolinguistic and sociocultural factors, such as Meyerhoff's (1992) research into the ethnic divide in NZE *eh* usage and Berland's (1997) dissertation on teenage usage of InTs in London. Despite the number of studies that categorize InT functions, such as Norrick's (1995) work on *hunh* and Avis's (1972), Gibson's (1977), and Gold's (2005) studies of Canadian *eh*, the range in methodology (corpus, interview corpus, and surveys/questionnaires, respectively) and reporting make their results incomparable. In other words, there are no descriptions of the frequency, range, and functions of InTs across varieties available to date.

Similarly, from a global English or teaching ESOL perspective, InTs have remained unexplored. Most ESOL texts devote at least one chapter to the correct use, formation, and meaning of canonical question tags. Yet while references to words such as *yeah*, *right*, and *okay* are made, it is extremely rare to see any mention of their use as InTs. A description of the InTs used in each variety may go toward such observations being included in English textbooks. This in turn should aid in intervarietal intelligibility. This study aims to describe InTs in NZE, British English (BrE), and Indian English (IndE) in detail with respect to the overall functions available and in particular the functions/meanings of a subset (*yeah*, *no*, *na*, and *eh*). The relative frequencies of InT functions within and across each variety will also be examined, and the four most frequent functions will be described in detail. The present study builds on previous research on InTs using English corpora, listing the functions of four InTs found in three varieties of English, as well as defining the most frequent four functions.[3] The purpose is to describe the variation in InT use and meanings across the three varieties.

1.2 Invariant tag definition

There have been several different terms and definitions applied to words like the four in question here. For example, Biber et al. (1999) call these words *response elicitors* because they aim to elicit a response (either verbal or gestural) from the listener (see also Holmes, 1982). However, such terminology is controversial in that responses to such items have not necessarily been found in corpus studies (e.g., Berland, 1997). The terminology used here, then, is *invariant tag* (henceforth InT), following Berland, with assumptions as follows:

InTs are question tags that do not change form, such as *Huh? Eh? Yeah? and Hey?* For example, in canonical question tags like *He hates math,* ***doesn't he****?* and *They don't hate math,* ***do they****?* there is a polarity change and subject agreement; in InTs, no such agreement issues arise, as we can see in *He hates math,* ***eh****?* and *They don't hate math,* ***eh****?*

Thus my definition is that an InT is a tag that is not a canonical question tag, in that there is no polarity issue and no subject agreement (e.g., *I can get them photocopied and send them out to people if that would be easiest,* ***eh****; And you approve of that,* ***huh****?*). InTs, as defined here, can elicit a response from the hearer or promote feedback or interaction in conversation, though this is not a requirement. They also offer the speaker's attitude (i.e., toward the hearer, the topic, themselves) beyond the propositional level of the utterance. Most importantly, an invariant tag is considered not to be exclusively a tag to an utterance—and thus found only in utterance-final position—but a tag to a concept or construction in the utterance, and therefore potentially occurring either utterance-finally, -initially, or -medially.

2. Methodology

Each variety under investigation (NZE, BrE, and IndE) was chosen as an example of a British-type English dialect. Thus BrE, the native and source variety, is compared to another native variety (NZE) and a native/lingua franca variety (IndE). These varieties also have the advantage of being both geographically distant in a Global English perspective and available in International Corpus of English (ICE) form, as ICE-GB, ICE-NZ, and ICE-IND. A subcorpus of private spoken dialogue (of 200,000 words)[4] was extracted from each corpus and loaded into Wordsmith 4 for the searches.

First, a set of four frequent tags, *yeah, eh, no,* and *na*, was chosen for in-depth functional analysis. The semantic and quantitative analyses were completed manually, using prior studies as background information rather than as a starting point for the meaning/function classifications. Given the lack of intonational mark-up or audio file availability,[5] all the analyses relied on a thorough reading of the full context for each concordance. Items that were not being used as tags were omitted from the description (with certain exceptions; see Results below), and any InT for which the meaning or function remained unclear was labeled as such. A full description and delineation of those similar functions, which were kept separate, is given in Columbus (in revision). Only functions of the four tags *yeah, na, no,* and *eh* are examined here; likewise, only the most frequent meanings are discussed.

2.1 Classification of InTs

As mentioned above, classifications were based solely on the written transcripts. Thus, all functions listed in the results have been based on close inspection and reading of the full context, and in some cases full text-file, of the concordance retrieved. Assigning a function or meaning to a tag involved reading the context several times and determining the speaker's intended meanings from what came before and after the example in the concordances. The categories used in the studies on tags conducted by Holmes (1982), Berland (1997), Stubbe and Holmes (1995), Norrick (1995), and Gold (2005) were used as background information

rather than as a template for the categorizations. Each potential InT meaning was marked for that concordance line, and the list of classifications was not assumed to be the same for each variety. Instead, each variety was coded separately, and category names which were highly similar between the varieties were conflated in the final analysis stage (e.g., Sarcasm/Humor). The multiple manual analyses allowed opportunities to re-evaluate the classifications made. To that end, most categories were validated and solidified through this process, being added to or excluded from the list given in Table 1. As with all semantic and discourse-based studies, however, it is true that a certain amount of subjectivity is unavoidable in assigning final meanings or functions. Despite this, many of the determined categories closely resembled those given in other tag studies, particularly those of Holmes (1982), Berland (1997), and Gold (2005). Thus it seems reasonable to assume that other researchers would have come upon very similar if not identical categorizations using the same dataset.

3. Results

3.1 Most frequent meanings

In Table 1 we see the full range of 20 functions for *yeah, na, no,* and *eh.* Of these, only 17 are true functions, the other 3 being unclear uses, filler uses (i.e., synonyms for *uh*), and the use of *eh* to mean 'pardon'. The latter two were included to highlight other common uses for these tag forms. Several categories are also combined, such as Prod/Turn Unit Extension (TUE) and Prod/Encouragement. Prod/TUE is a prodding function that serves to tell the listener that the speaker has completed his or her turn with the use of the InT (see Thompson and Couper-Kuhlen, 2005, for more on TUEs). Prod/Encouragement serves to use the InT as a minimal response marker, encouraging the speaker to continue their turn; the relatedness of these functions results in 16 separate meanings. Meanwhile Affirmation/Emphatic and Affirmation/Confirmation of Previous Statement are left separate to highlight the level of emphasis meant by the speaker. Examples for each function are given in Table 2. Note that the classifications were based on fuller contexts than are possible to print here.

Some tag occurrences could be assigned two separate functions. For example, Prod/Encouragement could combine with Softener, where the intention of the speaker is to gently push the interlocutor to continue. Classifications such as Emphatic or Affirmation could not be combined, as the tone of the intention stands alone. Likewise, Post Opinion/Statement is a function that adds no further intention to the meaning beyond "marking" an opinion or statement of fact. This contrasts with Confirmation Check, for example, in that this function clearly indicates the speaker's uncertainty regarding their statement. We now turn to an in-depth description of the four most frequent meanings for the InTs in this study: Confirmation Check, Emphatic, Narrative, and Post Opinion/Statement.

Table 1. Semantic classification distribution by variety.

Meaning		*yeah*		**InT** *eh*		*no*		*na*	
	IndE	NZE	BrE	IndE	NZE	BrE	IndE	NZE	IndE	Totals
affirmation: emphatic					7					7
affirm./conf. statement	3	1	4		47				11	66
checking question	1	8	3		21		24	1	9	67
comment on statement	3	2	5	1	53		3		10	77
confirmation check	2	10	13	1	75	2	112	2	35	252
empathetic			1		11					12
emphatic	5	7	8	2	103	3	31		28	187
exclamation/emphatic					2					2
narrative	18	20	27	1	81	1	62		59	269
new topic				2	4					6
offer/suggestion		1	3		15		5		6	30
pardon					22	4				26
post opinion/statement	9	4	8		116	2	56		18	213
prod/encouragement		18								18
prod/TUE					9	2				11
really?/check question		20								20
sarcasm/humor	2	2	3		14		2		5	28
softener					15					15
uh				1		2				3
unclear							3			3
Total	43	93	75	8	595	16	298	3	181	1312

Table 2. Concordance examples of each function.

affirmation: emphatic
B: Except when people get (laughs) old and then they always chop it off (laughs) A: Yeah I know B: (in high-pitched voice) They do *eh* (ICE-NZ [S1A-036#116–118])
affirmation/confirmation of previous statement
C: They don't have a rounded base A: *Na* they're shal- shallow base <,,> (ICE-IND [S1A-007#147–148])
checking question
M: Why is the light the light flashing on that F: Got to the end of the side? *No*? The light's flashing because M: Could've been a slow one? G: Probably it's just recording or something (ICE-NZ [S1A-006#186–190])
comment on previous statement
A: Yeah they've obviously taken on new staff because I noticed that the guy that used to work up here in Fairfield working at {the place in the Hutt} A: Oh do they? B: {They alternate} B: They're everywhere those two guys *yeah* (ICE-NZ [S1A-050#178–181])
empathetic
A: That's that's enough to keep you interested A: At least you're doing something *eh* (ICE-NZ [S1A-052#62–63])
emphatic
K: That's exactly what Kev said. And I said look um I think it's .. I sort of looked at it like that and I thought you know *yeah* why do they? They don't need to know as much about Maori because they're not like us. We're teachers, it's a bit different. (ICE-NZ [S1A-030#37–41])
exclamation/emphatic
A: And when we first got the letter from the energy board in Gisborne they wanted to pay us ten cents a year {1#<.>} for the (laughs) rent of it A: {2#(in high pitched voice) (unclear word) yes (word)} A: And then we said oh wasn't that a bit poor {3#you know paying} that much A: .... B: {2# Wow (in high pitched voice) ridiculous *eh*} (ICE-NZ [S1A-084#447–454])
narrative
N: Oh I've got all these books, heaps of them eh. (Laughs) All round here and (unclear word) R: Mm. N: So I make them available to my moko when he comes *yeah* and he just...And then... <,,> (ICE-NZ [S1A-080#299–304])
new topic
B: You should have listed some <,> something socially relevant and most in thing C: *Eh* do you think uh <,> the <,,> funds that they are collecting for the

earthquake is real worth of <,> (ICE-IND [S1A-056#47–48])

offer/suggestion
A: I don't know whether you want to
A: I can get them photocopied and send them out to people if that would be easiest *eh*
B: Yeah
A: And that will save you doing this ringing up (ICE-NZ [S1A-096#95–98])

pardon
A: Oh we can organise a boil-up eh
B: *Eh*? (ICE-NZ [S1A-087#212–213])

post opinion/statement
B: I'm an acquaintance of Phil Roberts
P: Oh yeah
P: He's pretty intelligent *eh*
B: Yeah ICE-NZ ([S1A-072#398–401])

prod/encouragement
F: But in her mind there's always going to be a family court judge deciding disputes, {custody} and stuff. And she said that in the c- in the course of doing her degree she's read some really crappy decisions
A: Mm
A: *Yeah*? (ICE-NZ [S1A-055#13–17])

prod/TUE
B: You know decent sized ones eh if it comes {to shitty} sizes
B: *Eh*
A: {Yeah}
A: Cos I wouldn't mind a few fish heads (ICE-NZ [S1A-087#202–204])

***really*?/checking question**
F: She wants to be a Family Court judge
A: *Yeah*? Get involved in all the family disputes that'd be quite hard
F: It's ... well sh- ... yeah I think so too (ICE-NZ [S1A-055#2–5])

sarcasm/humour
D: {(laughs) God you're nosy}. Don't believe this guy (laughs) Is there anything else you can nose through (laughs)
J: {no}
P: I think there's Jo's room actually *yeah*
D: (laughs)
J: There's nothing interesting there
P: No (ICE-NZ [S1A-073#322–329])

softener
S: I think if you look like you're taking the piss out of the whole thing <,> then it won't be appreciated very much *eh*
H: Yeah
S: But if you just show a humorous aspect of yourself with some cutting satire on a on a level (ICE-NZ [S1A-009#39–41])

'uh' filler use
B: You had to say one thing about <,,> your job and he mimicked that he he he did uhm *eh* demonstrate the mimic he demonstrated that in you know <,,> gave an indication of what that should sound like (ICE-GB [S1A-060#133])

3.1.1 Confirmation Check

This function seeks confirmation that the listener indeed shares the speaker's belief. Confirmation Check is consistent across all varieties in that it is the only function used for all four tags and it exists in each of the varieties of English examined here. In terms of meaning, it is perhaps the closest to canonical question tags of all the functions found here; that is, it aims to determine that the previous proposition is true, as canonical question tags such as *did they?* and *isn't it?* often do (though see Holmes, 1982**,** for the full range of functions for canonical tags). However, this function of verification also underlies other InT functions listed here. The difference between this function and other checking functions is the level of certainty on the behalf of the speaker. For instance, the Really?/Checking Question function listed in Table 1 is used when the speaker is surprised by the information he or she has just received. Conversely, the Checking Question function shows the speaker is less certain of the statement he or she has just made and appeals to the hearer to confirm. In example (1) of a Confirmation Check, below, we find *yeah* used in BrE to confirm Speaker A's almost certain beliefs regarding B's and C's wishes.

(1) ICE-GB (S1A-002#94–99)
A: So you you're both interested in performing <,> within this integrated <,> dance group *yeah*
C: Uhmuhm. Uhmuhm
A: That appeals to you both <,>
B: Uhm
C: Yes

3.1.2 Emphatic

This function is exactly as it is labeled: it emphasizes the propositional meaning intended by the speaker, making his or her attitude toward the statement more overt. This function can appear at the end of the utterance or within the utterance. It is found for *yeah, na, eh,* and IndE *no*. Example (2) illustrates a NZE use of *yeah* with this function. There is no semantic affirmation here with the use of *yeah* as far as the proposition is concerned; instead the incredulity of the following question clause is highlighted.

(2) ICE-NZ (S1A-030#37–41)
K: That's exactly what Kev said. And I said look um I think it's ... I sort of looked at it like that, and I thought you know, *yeah,* why do they? They

don't need to know as much about Maori because they're not like us. We're teachers. It's a bit different.

3.1.3 Narrative

The Narrative function appears to acknowledge the listener in the discourse but without meaning to elicit a response. It is a function that has also been found in previous tag studies, listed as *story-telling* by Gold (2005) in her study of Canadian *eh*, following Avis's (1972) and Gibson's (1977) *narrative* classification. The Narrative function could be labeled the opposite of a minimal response, in that rather than saying, "Keep talking, I'm listening," to the discourse partner, it says, "Keep listening, I'm (still) speaking." Indeed, Andersen (2001: 135) classifies such items as *non-turn-yielding tags*. This function most clearly illustrates the choice to label these *invariant tags* rather than *response elicitors*, since there is no response or other interruption desired by the speaker. Example (3) is an illustration of the speaker's use of *yeah* to show that the story is continuing, without emphasis or the expectation for a response, despite the fact that the speaker soon relinquishes his or her turn.

(3) ICE-NZ (S1A-080#299–304)
N: Oh I've got all these books, heaps of them eh. (Laughs) All round here and (unclear word).
R: Mm.
N: So I make them available to my moko when he comes *yeah* and he just ... And then ... <,,>
R: Yeah.

3.1.4 Post Opinion/Statement

This function is used to mark an opinion or statement made by the speaker and is speaker-centered in its meaning. Post Opinion/Statement tags add no further attitudinal information but instead signal that the hearer may now comment on the statement if he or she wishes. It is different from the Confirmation Check in that it is not clear that the speaker *knows* what the hearer thinks; the speaker simply makes his or her feelings known and adds a tag to make it less committal. In (4) below, Speaker A is not asking if his assumption is correct, but rather is making a claim and leaving the conversational space for the interlocutor to comment.

(4) ICE-NZ (S1A-072#398–401)
B: I'm an acquaintance of Phil Roberts.
P: Oh yeah.
P: He's pretty intelligent *eh.*
B: Yeah.

3.2 Meanings and preferred positions for the four selected tags

In Table 3, we see the total raw occurrences of *yeah, no, na,* and *eh*, with typical examples of concordances for each. Additionally, the positional preferences for the tags in each variety are given in Tables 4–7. A summary of the usage of each follows.

Table 3. Corpus frequencies and examples of the four detailed InTs in BrE, NZE, and IndE dialects.[6]

InT	Example	Frequency
eh	He's pretty intelligent *eh.* (ICE-NZ [S1A-072#400])	584
no	And does she get marks for it *no.* (ICE-NZ [S1A-036#339])	391
yeah	So we'll go and see what that's like *yeah.* (ICE NZ [S1A-035#23])	210
na	Air freight it takes about three days *na*? (ICE-IND [S1A-094#76])	181

3.2.1 *Yeah* functions and positions

Yeah occurred 210 times in the corpora and was the (nominally) preferred InT for BrE of the four tags with 74 raw occurrences. In BrE as well as IndE, it is often used for the Narrative function, at 47% and 36% respectively, as seen in Table 1. *Yeah* is also used for the Post Opinion/Statement and Emphatic functions. However, its use in NZE is somewhat different: while NZE uses *yeah* for Narrative, it is also used for the Prod/Encouragement and Really?/Check Question functions. In this it seems to be an extension of the basic BrE functions, although it shares the (low) use of Offer/Suggestion with BrE.

Table 4. Breakdown of *yeah* according to position (raw occurrences).

Position	Example	IndE	NZE	BrE	Total
nonclausal	*Yeah*?	0	41	0	41
initial	*Yeah* we have ladies bus	6	0	1	7
medial	When I was working *yeah* I was working with one organisation	13	9	22	44
final	just so she could put my hair up *yeah*	24	43	51	118
Total		43	93	74	210

From a positional perspective, as we can see in Table 4, the utterance-final positions are preferred across the varieties ($\chi^2 = 125.05$, $df = 3$, $p < 0.001$). Note that this supports the use of the term *tag*, despite InTs occurring in nonfinal positions. There are significant differences between the varieties, however, with respect to position (*Fisher's T-test*, $p < 0.001$). In NZE, *yeah* occurs more frequently than would be expected in the nonclausal position. On the other hand,

yeah is overrepresented in the clause-initial position in IndE, while BrE has fewer occurrences than would be expected in the nonclausal uses.

3.2.2 *No* functions and positions

In IndE, *no* is the preferred tag, forming 60% of the total of the four tags described here. It occurs almost twice as often as *na* in IndE. However, *no* is particularly rare in NZE (with only three occurrences and two meanings; see Table 1) and is completely absent in the BrE corpus. While this variation may possibly be due to sampling techniques, it seems more likely that NZE uses a different tag more frequently (see *eh* below), while BrE instead prefers other tags (such as the well-documented *innit* and *you know*) not investigated here.

Table 5 illustrates the positional preferences for *no*. Where these are concerned, *no* is clearly preferred in the final position ($\chi^2 = 603.93$, $df = 3$, $p < 0.001$). Though clause-medial tags are common, this position fails to reach significance as a preference.

Table 5. Breakdown of *no* according to position (raw occurrences).

Position	Example	IndE	NZE	BrE	Total
nonclausal	*No*?	0	2	0	2
initial	-	0	0	0	0
medial	We'll ask that person *no* that Sagar	90	0	0	90
final	and does she get marks for it *no*?	298	1	0	299
Total		388	3	0	391

3.2.3 *Na* functions and positions

Na is the only item of non-BrE origin that occurs frequently in any of the three subcorpora as a tag (see Table 1).[7] It occurs only in IndE and is *not* an alternative spelling of *no*. This tag is considered to have several origins, coming from Hindi and Urdu *na* as preverb negator, a short form of the *nahi* negator particle (Hamblin, 1984), and an Anglicized version of the Punjabi *naah* (Singh Malhotra, p.c.).[8] *Na* is almost a synonym for IndE *no* except for its use for Affirmation/Confirmation. It is interesting to note that this requires a certain amount of semantic bleaching given the origins of the tag.

Table 6. Breakdown of *na* according to position (raw occurrences).

Position	Example	IndE	NZE	BrE	Total
nonclausal	*Na*?	1	0	0	1
initial	*Na* they're shallow base	1	0	0	1
medial	Eighty *na* yes	74	0	0	74
final	But there is no use *na*	105	0	0	105
Total		181	0	0	181

In Table 6 we see that *na* occurs frequently in the clause-medial and clause-final position. Positive residuals are found for both the medial and final positions (χ^2 = 183.71, *df* = 3, $p < 0.001$). Notice that this is another distinction between the use of *no* and *na* as InTs in IndE.

3.2.4 *Eh* functions and positions

Of the four tags described here, *eh* is the most frequent and the most flexible (see Table 1). *Eh* as a tag is the most preferred in NZE—it forms 85% of all the NZE InTs. Conversely, *eh* has comparatively rare usage in BrE (16 raw occurrences) and IndE (8 occurrences), with 7 and 6 functions respectively compared to 16 in NZE. Of these functions in BrE and IndE, *eh* is often used for filler (viz., *uh*) and Pardon uses, which are not tag functions per se. Interestingly, the only use of a tag for New Topic is with IndE and NZE *eh*. Most frequently, however, NZE *eh* is used for Emphatic, Narrative, Confirmation Check, and Post Opinion/Statement functions, contributing to their status as the most common functions overall. NZE *eh* is not used as a filler, for Really?/Checking Question (this is the domain of *yeah* in NZE), or for Prod/Encouragement.

In terms of positional preferences (Table 7), *eh* is overrepresented as an utterance-final tag (χ^2 = 803.19, *df* = 3, $p < 0.001$). Indeed, *eh* is overrepresented in the NZE data (χ^2 = 1028.6, *df* = 2, $p < 0.001$). There is also a correlation between Position and Variety (Fisher's T-test, $p < 0.001$). This is due first to the overrepresentation of *eh* in utterance-initial uses, second to the overrepresentation of *eh* in BrE non-clausal uses, and finally because of the underrepresentation of *eh* in IndE utterance-final position. Here it is clear that while NZE *eh* prefers final positions, BrE and IndE do *not* prefer final uses. The results here, and for *yeah, no,* and *na*, mean that tags are not categorically utterance-final.

Table 7. Breakdown of *eh* according to position (raw occurrences).

Position	Example	IndE	NZE	BrE	Total
nonclausal	*Eh*?	0	26	5	31
initial	*Eh* do you think uh the funds oh I've got all these books	6	7	1	14
medial	heaps of them *eh* all round here…	2	98	2	102
final	They do *eh*	0	429	8	437
Total		8	560	16	584

4. Discussion

The results above confirm that while InTs are similar in many respects across the NZE, BrE, and IndE varieties, they are not completely interchangeable. The 17 meanings found in total for the four tags are not distributed evenly across the English dialects. While *na* is available only in IndE, it shares almost all its

functions with IndE *no*, with the exception of the Affirmation/Confirmation function. NZE *no*, on the other hand, is both very uncommon and restricted to only the most general tag functions—Checking Question and Confirmation Check. Likewise, where NZE *eh* has 16 functions available to it, IndE *na* has only 9, and BrE has no one tag that is available for such a range of meanings. Furthermore, *yeah* is the most common InT across the three corpora, being used to a relatively equal extent in BrE, IndE, and NZE. Its functions, however, are split into two groupings, with similar types and frequencies for functions in BrE and IndE, and an extra use available to NZE speakers.

There are further examples of BrE and IndE patterning together compared to NZE with respect to these tags. For example, BrE and IndE InTs have fewer functions available to them than the NZE InTs (10 and 11 respectively, but 17 for NZE). Similarly, *yeah* use for Affirmation of Previous Statement and Post Opinion/Statement is more frequent in BrE and IndE than in NZE. However, for the majority of tags and functions, NZE use is much more frequent, and IndE and BrE have comparable, lower occurrences for each. This is the case for Emphatic and Confirmation Check use of *eh*, Narrative use of *eh*, and Prod/Encouragement or Prod/TUE across all InTs. Also, the use of *eh* as a pause filler (meaning 'uh') is restricted to BrE and IndE. Thus even the smallest of words shared across these varieties can distinguish two groups of tag users—BrE and IndE speakers in one group and NZE speakers in the other. This is further supported by the range and raw occurrences of all InTs in BrE and IndE when compared to NZE (Columbus, in revision).

Maybe most interesting are the functional differences between the InTs. While alternatives between varieties and dialects are immediately marked in terms of single-word vocabulary (viz., *trash—garbage* or *rubbish*, or *beanie—hat* or *toque*), the use of tags and potential attitudinal meanings for them are not obvious. It is conceivable that NZE speakers using *yeah* in a BrE setting may find themselves using it for a function that does not exist in the British sense, and vice versa. For example, a NZE speaker might make a suggestion using *eh*, such as *I can get them photocopied and send them out to people, that would be easiest,* ***eh***. Yet it is not certain that a BrE hearer would understand its function, given the BrE equivalent is *yeah* (e.g., *Yeah but w uh we can come round* ***yeah*** *and ...*). Similarly, if a BrE speaker used *yeah* as a Checking Question or Confirmation Check, which also both occur in NZE, the higher frequency of *yeah* as a Really?/Check Question in NZE may predispose the NZE hearer to interpret the BrE speaker's statement as showing surprise. Additionally, the virtual nonuse of *no* as an InT in NZE and BrE would almost certainly lead to confusion as to the speaker's attitude when used by an IndE speaker. These differences in potential functions are perhaps clearest with respect to the extremes of politeness. That is, BrE and IndE have no InTs that can be used for softener/hedging functions, and NZE use of *eh* for this purpose may not be interpreted as such (e.g., *it's best to be polite* ***eh*** *otherwise they they start asking questions* ***eh*** *if they if they see if they um see there's something um fishy about it*). The rare use of sarcasm and humor in BrE and IndE may likewise cause confusion or worse—offense—as to the NZE

speaker's intentions in using *eh* for sarcasm (e.g., *but you were the one by yourself in the corner* ***eh***). Thus speakers of different English varieties, not to mention nonnative speakers, may find themselves at a loss to determine the exact intent of a speaker using an InT from another variety.

At a Global English level, it is worthwhile noting that BrE and North American English are often used as standards for teaching ESOL. The differences found between tags here suggest that problems may occur for even proficient nonnative speakers of English at the dialectal level. We have seen here that the subtle inferences in the use of a question tag such as *eh* in NZE cannot be seen by mere analysis of form. Neither is the difference between the NZE and BrE *eh* visible through intuition and morphological studies alone. When we consider that ESOL pedagogy rarely refers to InTs, despite a determined focus on canonical question tags, it becomes clear that a level of meaning above the propositional level may never be accessible to the nonnative user of English. What would happen in both intervarietal and Global English use of InTs is unclear; InTs do not change propositional meaning, and so complete miscommunication is unlikely. However, given that such minimal changes as intonation can alter a hearer's perception in a conversation (e.g., Gumperz's 1982 study on InE in a British environment), subtle meaning differences in speaker attitude may indeed cause problems at the discourse level.

The variation in usage of conversational techniques is also of interest. TUEs occur cross-linguistically in different formats. They essentially act to continue a turn that seemed completed, typically for the purpose of demonstrating that the speaker's turn is over. The conversational floor is then left open for the interlocutor (Lindström, 2006). Turn-taking and turn-holding functions such as Prod/TUE and Narrative are unevenly spread across the InTs and varieties as described here. As mentioned above, turn-taking and turn-holding uses of tags occur mostly in the NZE data. Specifically, BrE uses *eh* as a Prod/TUE, while NZE uses *eh* for Prod/TUE and *yeah* for Prod/Encouragement. It is probable that other methods must be used for these functions in the varieties, as they exist cross-linguistically (Thompson and Couper-Kuhlen, 2005). Since the tags here are infrequently used for this purpose, NZE, BrE, and particularly IndE must have other TUE particles or phrases available to their speakers. Whether these are other discourse markers (e.g., *ellerhur* 'right' or *tyckerja(g)* 'I think' in Swedish; Lindström, 2006), common phrases (e.g., *d'you know* in English; Thompson and Couper-Kuhlen, 2005) or noun phrases (as in Japanese; Thompson and Couper-Kuhlen, 2005), they remain to be seen. More interesting, however, is the possibility that other varieties may choose InTs for this function over other discourse markers, as NZE does with respect to *yeah*. As such, a full investigation of cross-varietal turn-taking and turn-holding could identify differences in discourse style for these English dialects based on InT choice and usage.

Finally, the results show that the default position for InTs is the utterance-final position, regardless of variety. This echoes the traditional placement of canonical question tags. Yet the use of InTs in utterance-medial position also reaches significance for *na*. Utterance-initial and nonclausal positions fail to

reach significance, but their overrepresentation for some tags in some varieties highlights the fact that the final position is not the only place available for tags. Thus the four InTs investigated here act more often like question tags in positioning but are like other discourse particles in being able to occur initially, medially, and, unlike some other discourse markers (such as *like* in most varieties; see Miller and Weinert, 1995), as complete utterances on their own.

5. Conclusion

We have seen that InTs are found across the three English varieties investigated here, but like single vocabulary items, the tags are not the same in each variety. While BrE, IndE, and NZE share a certain number of tags and tag functions, no single tag or tag set in a single variety is entirely matched across the range of meanings. Similarly, the frequency of tags is not equivalent across the varieties, with NZE and IndE having distinct preferences (*eh* and *no*), and BrE having no clear preference within the four tags selected. The implications for such differences are not certain. However, the likelihood of misunderstandings, though not communication breakdown, is high when one considers the subtle mismatches in attitudes and expectations in using a mother tongue tag in another variety. This means that mutual intelligibility of English is not entirely possible at the discourse level, a fact that is remarkable when the propositional meaning remains the same. Given these differences, and those of varietal position preference within an utterance, it seems reasonable that texts and dictionaries for English for Speakers of Other Languages, and in particular those for Academic and Business English, include at least some reference to these items, if only to list them and highlight that usage differs between varieties.

Furthermore, other curious facts come to light when InTs are compared in this way. As mentioned above, in many ways BrE and IndE pattern closely together, leaving NZE InTs with alternative choices in possible tags and functions. This could be indicative of a closer varietal relationship or perhaps an indication of the prestige/aim dialect in IndE (see Kachru, 1994; Schneider, 2007) and of extension of and divergence from the origin variety in NZE. It may be interesting to see if other inner-circle varieties of English have similar extension/divergence patterns as found in NZE InTs. Additionally, the positional preferences for InTs in these varieties show homogeneity for the most part. There are, though, some differences between these three English varieties in terms of positional spread. For example, NZE *yeah* has a high one-word-utterance usage with relatively little utterance-medial use, and NZE *eh* is used in one-word, medial, and final positions. Conversely, while final position is most preferred for the four InTs described here, varieties do not always select the same positions for each of the tags, such as with medial *na* in IndE. Overall then, there are more differences than similarities between IndE, BrE, and NZE InTs. And if such differences exist in just three varieties of British-origin English, what would be

uncovered if other varieties, including Northern American varieties, were to be compared in such a way? This question must remain one for future research.

6. Notes

* I would like to thank John Newman at the University of Alberta and the reviewers for their valuable feedback on this paper. Thanks are also due to the audience for this paper at AACL 2008, as well as those at presentations of papers building on this study given at ICAME 2007 and Methods XIII 2008, for their comments. All errors, of course, are my own.

1. I would like to thank John Newman at the University of Alberta and the reviewers for their valuable feedback on this paper. Thanks are also due to the audience for this paper at AACL 2008, as well as those at presentations of papers building on this study given at ICAME 2007 and Methods XIII 2008, for their comments. All errors, of course, are my own.
2. Such as Irish English *like* being used as a tag (Kallen p.c.) in contrast to Canadian English *like* being used as a quotative (Tagliamonte and D'Arcy, 2004).
3. For a full description of all InT functions, see Columbus (in revision); for frequencies and comparison to further varieties, see Columbus (forthcoming) and Columbus (submitted).
4. The files utilized from each ICE corpus were the Private Spoken Dialogue files: Face-to-face Conversation texts S1A-001–S1A-090, and Telephone Conversation texts S1A-091–S1A-100. These represent 200,000 words per corpus of the total 600,000-word spoken subcorpus.
5. The ICE-GB audio corpus had not yet been released at the time of the study.
6. I would like to thank the editors for their aid in the statistical analysis.
7. The preliminary search of ICE-NZ included the Maori terms *ae* (yes) and *kao* (no). These were not found to be used as tags within English sentences, however.
8. Lawler (n.d.) states that some parts of India are reported to use *ah* for the same function. However, this form was not attested in this corpus.

References

Andersen, G. (1998), "Are tag questions questions? Evidence from spoken data," paper presented at the 19th ICAME Conference, Belfast, United Kingdom.

Avis, W. (1972), "So *eh?* is Canadian, eh?" *Canadian journal of linguistics*, 17: 89–105.

Berland, U. (1997), *Invariant tags: pragmatic functions of* innit, okay, right, *and* yeah *in London teenage conversations*. Unpublished master's thesis, University of Bergen, Norway.

Biber, D., S. Johansson, G. Leech, S. Conrad, and E. Finegan (1999), *Longman grammar of spoken and written English.* Harlow: Longman.

Columbus, G. (forthcoming), "A corpus-based analysis of invariant tags in five varieties of English," in: A. Renouf and A. Kehoe (eds.) *Corpus linguistics reassessed. Papers from the 28th international conference on English language research on computerized corpora (ICAME 28).*

Columbus, G. (in revision), "A comparative analysis of invariant tags in three varieties of English," *English world wide.*

Columbus, G. (submitted), "'Nice day, eh?': Canadian and New Zealand *eh* compared," *Proceedings of methods in variation XIII 2008.*

Gibson, D. (1977), "Eight types of 'eh,'" *Sociolinguistics newsletter,* 8(1): 30–31.

Gold, E. (2005), "Canadian *Eh?*: a survey of contemporary use*," Proceedings of the 2004 annual conference of the Canadian linguistic association*.

Gumperz, J. J. (1982), *Discourse strategies*. Cambridge: Cambridge University Press.

Hamblin, C. (1984), *Languages of Asia and the Pacific.* London/Sydney: Angus and Robertson.Holmes, J. (1982), "The functions of tag questions," *English language research journal,* 3: 40–65.

Kachru, B. B. (1994), "English in South Asia," in: R. Burchfield (ed.) *The Cambridge history of the English language, Vol. V*: *English in Britain and overseas: origins and development*. Cambridge: Cambridge University Press.Lawler, J. (n.d.), "Indian English: grammar tweaks". Online at <http://en.wikipedia.org/wiki/Indian_English>.

Lindström, J. (2006), "Grammar in the service of interaction: exploring turn organization in Swedish," *Research on language and social interaction,* 39(1): 81–117.

Meyerhoff, M. (1992), "'We've all got to go one day, eh?': powerlessness and solidarity in the functions of a New Zealand tag," in: K. Hall, M. Bucholtz, and B. Moonwomon (eds.) *Locating power: proceedings of the 2nd annual Berkeley women and language conference.* Berkeley, CA: Berkeley Women and Language Group, 409-419.

Miller, J., and R. Weinert (1995), "The function of LIKE in dialogue," *Journal of pragmatics*, 23: 365–393.

Norrick, N. R. (1995), "*Hunh*-tags and evidentiality in conversation," *Journal of pragmatics,* 23: 687–692.

Schiffrin, D. (1987), *Discourse markers.* Cambridge: Cambridge University Press.

Schneider, E. W. (2007), *Postcolonial English: varieties around the world.* Cambridge: Cambridge University Press.

Stubbe, M., and J. Holmes (1995), "*You know*, *eh,* and other exasperating 'expressions': an analysis of social and stylistic variation in the use of pragmatic devices in a sample of New Zealand English," *Language and communication,* 15: 63–88.

Tagliamonte, S. A., and A. D'Arcy (2004), "'He's like, she's like': The quotative system in Canadian youth," *Journal of sociolinguistics,* 8(4): 493–514.

Thompson, S., and E. Couper-Kuhlen (2005), "The clause as a locus of grammar and interaction," *Discourse studies,* 7(4–5): 481–505.

Good nouns, bad nouns: what the corpus says and what native speakers think

*Philip Dilts**

Department of Linguistics, University of Alberta

Abstract

Many researchers have found that some words or constructions tend to co-occur with words representing a positive or negative semantic nuance, demonstrating that these words have a certain semantic preference *(e.g., the negative preference of* cause *in Stubbs, 1995). Other researchers have explored the negative or positive associations of words taken out of context, their* semantic orientation *(Osgood et al., 1957; Turney and Littman, 2003). In this paper, we investigate how well a word's semantic orientation correlates with its semantic preference. We use the quantitative method developed by Dilts and Newman (2006) to measure how strongly a large number of nouns in the British National Corpus prefer to collocate with positive or negative orientation adjectives. We then compare each noun's semantic preference to its rating for 'pleasure' from the Affective Norms for English Words (ANEW) (Bradley and Lang, 1999), an established psychological measure of a word's semantic orientation. We find a surprisingly large number of nouns with negative semantic orientation but positive semantic preference: that is, 'bad' nouns preferring to collocate with 'good' adjectives. By contrast, only a small number of 'good' nouns attracted more 'bad' adjectives. Our results suggest an interesting mismatch in the way nouns are modified: while 'good' nouns attract primarily positive adjectives (further reinforcing their semantic orientation), 'bad' nouns attract both negative (reinforcing) and positive (qualifying) adjectives that have a greater transformative effect on the semantics of the noun.*

1. Introduction

Affective and evaluative language is the subject of vibrant study in several areas of current linguistic research, from corpus linguistics (Bednarek, 2008; Xiao and McEnery, 2006) to computational linguistics (Wilson et al., 2005) to psycholinguistics (Wurm, 2007). The present work looks at the interaction between two concepts developed in these areas: semantic preference and semantic orientation. In particular, we ask, "Do words that evoke positive feelings or represent positive evaluations (i.e., words with positive semantic orientations) tend to co-occur with (i.e., show a semantic preference for) other words with similar affective properties?" That is, do words prefer to collocate with words that reinforce their semantic properties or with words that transform them? Do the preferences differ depending on whether 'bad' or 'good' words are under consideration? This paper attempts to advance the discussion on each of these questions by contributing to the ongoing development of empirical and systematic ways in which they are being addressed.

2. Previous studies

This section outlines relevant previous work on semantic orientation, semantic preference, and the related notion of semantic prosody. Working definitions for the key terms, *semantic preference* and *semantic orientation,* are also provided, since the terms are not always used consistently across studies or authors.

2.1 Semantic orientation

In the present study, the semantic orientation of a word is defined by how good or bad people feel when presented with that word. A word like *fun,* for example, is consistently rated as pleasant by English speakers and is said to have a positive semantic orientation, while a word like *jail* is said to have a negative semantic orientation. Hatzivassiloglou and McKeown (1997) first used the term *semantic orientation* in their attempt to automatically discover adjectives representing positive or negative evaluations—that is, adjectives with good or bad semantic orientations. Systematic investigations of positive and negative language, however, go back much further. Osgood et al. (1957), for example, discovered that semantic orientation (their "evaluative factor" of meaning) is not only psychologically real, but also the most useful factor in predicting how people will judge objects or concepts.

The remainder of this section provides a brief overview of previous studies of semantic orientation. First, psychological studies of semantic orientation are described. Next, some computational studies aimed at automatically extracting the semantic orientation of several words from corpora are reviewed.

2.1.1 Psychological measures of semantic orientation

In their classic study of the measurement of meaning, Osgood et al. set out to discover the dimensions of semantic space represented in the mind. They selected 138 pairs of antonymous adjectives representing a wide variety of meanings. They first removed 62 of these pairs that were judged by participants to be similar in meaning to one or more of the other 76 pairs. A second set of participants then rated each member of a set of diverse concepts (*sin*, *opera*, *Richard Nixon*) on seven-point scales defined by each of these adjective pairs. By using factor analysis, Osgood et al. determined a small set of axes of meaning that described their subjects' ratings. The most prominent axis was tightly correlated with the scale from 'good' to 'bad.' That is, their participants rated concepts based primarily on their semantic orientation.

Bradley and Lang (1999) elicited affective ratings along three of Osgood et al.'s (1957) axes for over 1,000 English words. To measure semantic orientation, their participants rated how pleasant or unpleasant each word made them feel on a nine-point scale. The authors report the mean and standard deviation of these ratings as part of the Affective Norms of English Words (ANEW).

Wilson et al. (2005) trained researchers to annotate the semantic orientation of subjective expressions in the Multiperspective Question-Answering project corpus (MPQA) (Wiebe et al., 2005). Annotators were asked to decide whether almost 16,000 such expressions, which range from single words to complete utterances, were used to indicate positive or negative private states (opinions, emotions, etc., held by the author) when taken in context. The MPQA corpus can thus be seen as containing a corpus of semantic-orientation-tagged linguistic units analogous to ANEW, but with two crucial differences: In the MPQA, a word or expression's semantic orientation can vary depending on its context, while ANEW semantic orientations are elicited for isolated words. Second, the MPQA semantic orientation ratings appear to have been provided by a single annotator, while the ANEW ratings represent an average of responses from several participants. Nonetheless, Wilson et al. (2005) report agreement between a pair of annotators of over 80% for a subset of their corpus (447 expressions).

Recent studies using modern psycholinguistic techniques provide evidence both for the importance of semantic orientation in lexical processing (e.g., Wurm et al., 2004) and for a refinement of the concept. Wurm and Vakoch (2000), for example, have argued that the good-bad evaluative axis of semantic orientation can be profitably divided into two independently variable axes: one for danger and the other for usefulness. Wurm (2007) conducted a thorough psycholinguistic study and determined that danger and usefulness have separate robust effects on lexical decision and word naming times, and explain as much variance as the axes developed in Osgood et al. (1957), including semantic orientation.

2.1.2 Automatic extraction of semantic orientation from corpora

Hatzivassiloglou and McKeown (1997) attempted to automatically classify English adjectives according to their semantic orientations. The authors extracted adjectives in conjunctions (e.g., *slow but steady, fast and loose*) from corpora and grouped them into two categories based on the words with which they were conjoined. Words in the group with higher average word frequency were classified as positive, while words in the group with lower average frequency were classified as negative.

Several studies have attempted to automatically determine the semantic orientation of a word by considering the proximity to or prevalence of neighboring words with certain known (or assumed) semantic orientations. Baron and Hirst (2003) provide an unpublished, preliminary note on a potential research program of this type. The authors propose to estimate a word's semantic orientation by counting the percentage of its neighborhood that consists of words with known semantic orientation. The authors plan to use the General Inquirer lexicon (GI) (Stone, Dunphy, Smith, and Ogilvie 1966), in which several words are tagged as *positive* or *negative*. The authors intend to use this method to investigate the orientation of words in the British National Corpus (BNC).

Turney and Littman (2003) conducted an extensive study of automatic semantic orientation extraction. They first chose seven pairs of adjectives that

they believed to represent positive or negative evaluations in any context (e.g., *excellent* or *poor*). The authors could then assign a target word a semantic orientation based on how many of these known-orientation seed words appeared in a neighborhood around it in a corpus.

Turney and Littman tested several variations of their method. They used three different corpora (one of which consisted of all English-language web pages indexed by the AltaVista search engine), two different information-theoretic measures—Pointwise Mutual Information and Latent Semantic Analysis—to determine how strongly the seed words influenced the orientation score, and neighborhoods ranging from two words to 1,000 words in size. To measure the effectiveness of each technique, the authors used each combination of parameters to determine the semantic orientation of both all the items in the GI lexicon tagged as *positive* or *negative* and all the adjectives whose semantic orientation was determined in Hatzivassiloglou and McKeown (1997). Their most effective parameter settings achieved an accuracy of 82.8% on these sets.

2.2 Semantic preference and semantic prosody

The notion of semantic prosody is not new (its earliest form can be found in Sinclair, 1987), but it has been of considerable interest lately (e.g. Partington, 2004; Whitsitt, 2005; Xiao and McEnery, 2006). Despite this interest, there is little agreement on a definition of semantic prosody to date.

Several authors agree on what represents an example of semantic prosody: the construction *set in* tends to appear with subjects that represent negative or unpleasant concepts. For example, in the BNC we find the sentence, *An incredible arrogance and corruption and decadence set in*. Sinclair (1987) observed this fact and pointed out that this tendency is not readily available to native speaker introspection. Several authors have used this example to explain their perspectives on semantic prosody (Louw, 1993; Sinclair, 1996; Partington, 2004; Whitsitt, 2005).

Still, these studies differ in which aspect of constructions like *set in* they consider key in explaining why it exemplifies semantic prosody. In his initial definition, Louw (1993) emphasized that *set in* has a negative semantic prosody because by choosing the words *set in*, a speaker is making a pragmatic decision to indicate a negative attitude toward a subject. Stubbs (2001) also regards the pragmatic nature of such prosodies as central to their definition, even calling them "discourse prosodies" rather than semantic prosodies. Whitsitt (2005), on the other hand, thinks that this emphasis on speaker intention is antithetical to what he considers the original definition of semantic prosody. Citing Louw (1993), he writes that semantic prosody is a process of diachronic language change in which the frequent production (and perception) of negative subjects in instances of the *set in* construction has led to the negativization of that construction over time—that is, to its negative semantic prosody.

Definitions of semantic preference are more uniform. In semantic preference studies, researchers show that some word forms, or lemmas, tend to co-occur primarily with words that fall into a restricted set of semantic categories.

For example, the word *undergo* is typically followed by verbs representing change or training and testing, and preceded by words that indicate involuntariness (Stubbs, 2001). Thus, *undergo* has a semantic preference for verbs denoting change.

To describe the difference between semantic preference and semantic prosody, Stubbs notes that semantic prosody usually manifests itself at a higher level of abstraction and that semantic preferences are more restricted in size and semantic category. To exemplify the difference, he points out that the preponderance of words representing change that follow *undergo* and of words representing involuntariness that precede it represent interesting semantic preferences. The combination of these preferences, coupled with the notion that change and involuntariness are considered undesirable, lends a more general air of negativity to constructions involving *undergo*. This general negativity is the *undergo* construction's negative semantic prosody.

Stubbs (2001) and Partington (2004) both emphasize broadness or generality as a key characteristic of semantic prosody. In practice, though, this generality often takes a very specific form: a construction's semantic prosody is almost always phrased in terms of positive or negative evaluation. In their summary of ten semantic prosody studies, for example, Xiao and McEnery (2006) find that 19 constructions were shown to have a negative semantic prosody, while three constructions were shown to have a positive semantic prosody. There were no studies in which a construction was discovered to have, say, an active or passive semantic prosody, though this activity level was Osgood et al.'s (1957) second-most prominent axis of meaning.

Dilts and Newman (2006) attempt to bring more methodological consistency to the study of semantic prosody by allowing only empirically judged positive and negative words in a fixed-size search window to count toward a word's semantic prosody. As a result of these increased restrictions, however, their study falls on the border between semantic prosody and semantic preference. Like semantic preference studies, they consider the semantic category of the words in a very specific collocation frame. Like semantic prosody studies, on the other hand, the semantic category on which they focus is positivity or negativity—that is, semantic orientation. In the follow-up study described below, Dilts and Newman themselves take the position that their 2006 study represents an investigation of semantic preference.

Bednarek (2008) would likely argue that the subject of Dilts and Newman (2006) (and indeed, the subject of the current work) is positive or negative collocational semantic preference, which can sometimes lead to a positive or negative semantic prosody that we are unable to measure objectively. Bednarek (2008) carefully reconstructs and differentiates the terms *semantic preference* and *semantic prosody* in a way that clarifies and is consistent with seemingly contradictory existing works on the subjects. It is an invaluable resource for any future study.

2.3 Comparing semantic orientation to semantic preference

Dilts and Newman (2008) is a pilot study designed to investigate the relationship between semantic orientation and semantic preference. The authors presented the 50 nouns with the most positive or negative semantic preferences in the BNC (as identified in Dilts and Newman, 2006) to five native English speakers and asked them to decide whether the nouns represent good, bad, or neutral concepts. A simple majority agreement was accepted as enough to decide on each noun's semantic orientation. The results are shown in Table 1 below; values are in nouns, with accompanying percentages rounded to the nearest percent.

Table 1. Comparison of native-speaker ratings of nouns' semantic orientation and their collocational preference for good or bad adjectives. Adapted from Dilts and Newman, 2008.

Semantic preference	**Semantic orientation**	
	bad	good
positive	1 (3%)	12 (34%)
negative	12 (34%)	10 (29%)

Their results showed that nouns of both orientations often exhibited semantic preferences for reinforcing adjectives: good nouns attracted good adjectives, and bad nouns attracted bad adjectives. However, a relatively large number of good nouns preferred to collocate with bad adjectives, while almost no bad nouns preferred to collocate with good adjectives. They concluded that positive nouns appear to allow adjectives that transform their semantic orientation, while negative nouns do not.

3. Present study

The present study attempts to investigate the relationship between collocational semantic preferences and experimentally determined semantic orientations on a larger scale than Dilts and Newman (2006). Instead of using a small set of native speakers to judge semantic orientation, we use the more thoroughly studied semantic orientation measures provided by ANEW. Instead of considering only the 50 nouns with the strongest collocational preferences from Dilts and Newman (2006), we consider the set of over 700 nouns that have both semantic preference ratings calculated by Dilts and Newman (2006) and semantic orientation ratings established in ANEW.

3.1 Methods

We calculate the semantic preference for nouns in the BNC using the method presented in Dilts and Newman (2006), assigning preference values based on the relative proportion of good or bad adjectives immediately preceding the target noun. The following section outlines how these adjectives were chosen for the current study.

3.1.1 Adjective pairs

The metric developed by Dilts and Newman (2006) requires a set of pairs of seed adjectives with known semantic orientations. Each of these pairs can be weighted to have a larger or smaller effect on a collocating noun's semantic preference, depending on how confident the researcher is about the pair's semantic orientation, for example, or on how strongly the adjectives in the pair are oriented.

Osgood et al. (1957) established evaluative factor loadings (i.e., degrees of positive or negative semantic orientations) for 76 pairs of adjectives. These loadings were used as semantic orientation weights. The researchers also listed 62 adjective pairs that their participants found to be similar in meaning to at least one of the 76 adjective pairs with known orientation. They excluded these adjective pairs from their factor analysis, and as a result, no factor loadings have been empirically established. Consider the adjective pair *safe-dangerous*, for example. Osgood et al. do not provide an empirical measure of how well this pair of adjectives aligns with their good-bad axis, since it was considered too close to good-bad to be useful in their endeavor.

Should these adjectives be included in the analysis? If so, what weight should they be assigned? The choice involves a tradeoff between our confidence in the validity of the results and our confidence in their generality. If none of the 62 unmeasured adjectives are included, the number of adjectives used to make claims about the semantic orientation of a given noun is more limited. If a noun has a strong, idiosyncratic collocational preference for one of the few adjectives included in the seed adjective set, its semantic preference measure could be artificially inflated. This adverse effect should be less of an issue given a larger set of adjectives. In other words, the larger the set of seed adjectives, the more generalizable the results.

Using poor examples of positive or negative orientation as seed adjectives has its own pitfalls. In the best case, these adjectives will add noise to the calculations, decreasing the clarity and usefulness of the results. In the worst case, including systematically poor choices of seed adjectives will result in a systematic skewing of the calculations, leading to unfounded confidence in inaccurate results.

In the current work we perform and compare three separate analyses. The analyses differ in how the 62 adjective pairs without measured factor loadings are weighted. In what we call the *empirical* analysis, these adjective pairs are simply excluded from the calculations. In the *full-strength* analysis, the unmeasured adjective pairs are assigned the weight of the adjective pairs they were considered to be similar to in meaning. *Safe-dangerous*, for example, was given a weight of 1, since it was considered similar to the *good-bad* pair that has a factor loading of 1.

A third, *half-strength* analysis was also conducted. In this compromise analysis, the 62 uncertain adjective pairs were included, but with half the weight of their empirically measured counterparts. This allows these adjectives to have

some say in deciding a noun's semantic preference, but operationalizes our lack of confidence in their semantic orientation.

The full-strength and half-strength analyses found semantic preferences for 31,119 nouns, and the *empirical* analysis found semantic preferences for 27,428 nouns, all from the BNC. We then searched ANEW for pleasantness ratings for as many of these nouns as possible, and found more than 700 nouns in each case. We then compared the collocational preferences to the semantic orientations, as described in the following section.

3.2 Results

More than 700 of the nouns whose collocational semantic preferences we calculated had ANEW semantic orientation ratings. The relationship can be imagined as a two-dimensional space defined with semantic orientation as one dimension and semantic preference as another. Each word can then be represented as a point in this space, as in Figures 1 to 3 below. Each word is represented by either a word or a dot: words (*home, killer, joke,* etc.) represent words that are included in Dilts and Newman (2008), while dots represent words that are not. Lines at $y = 0$ and $x = 5$ represent neutral values. Curved lines are Lowess smoothers.

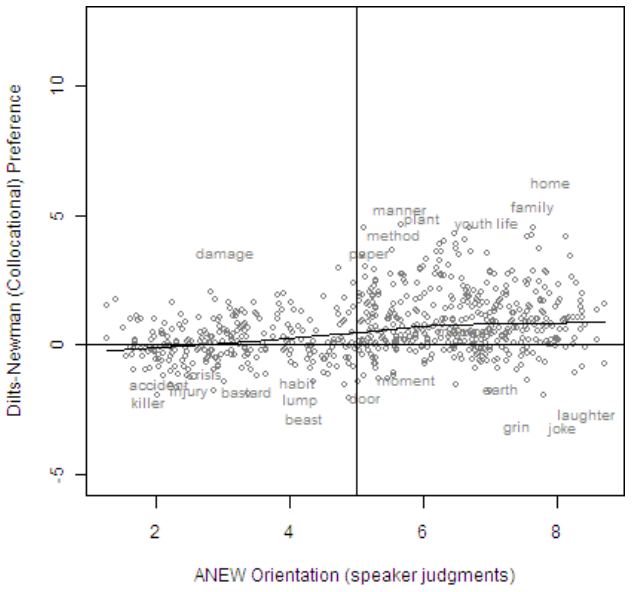


Figure 1. Semantic orientation vs. semantic preference for English nouns (empirical seed adjectives).

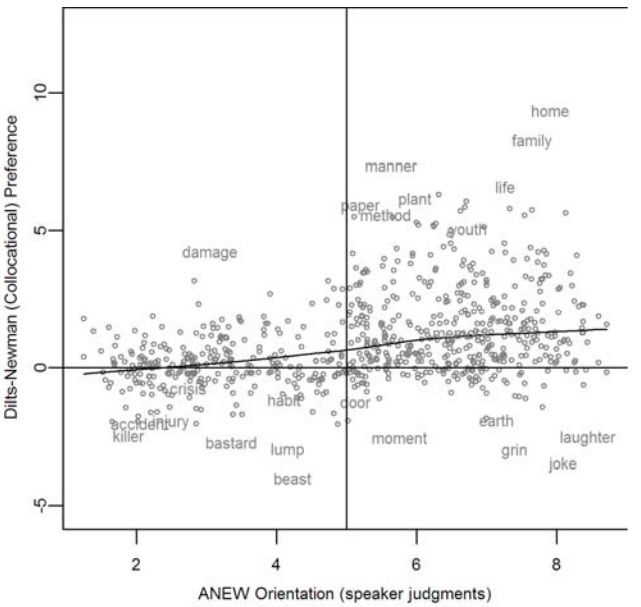


Figure 2. Semantic orientation vs. semantic preference for English nouns (half-strength seed adjectives)

Words falling to the right of each vertical line were rated as pleasant by ANEW raters, while words falling to the left were rated as unpleasant. Words above each horizontal line preferred to collocate with good adjectives, while words below the line preferred bad adjectives. Randomly chosen examples from each of the four quadrants from Figure 1 are given below.

(1) Positive preference, negative orientation nouns (top-left):
fire, prick, rat, razor, avalanche
(2) Positive preference, positive orientation nouns (top-right):
quiet, appliance, poetry, wish, key
(3) Negative preference, negative orientation nouns (bottom-left):
trouble, quarrel, debt, punishment, burial
(4) Negative preference, positive orientation nouns (bottom-right):
idol, child, sphere, father, kindness

These four categories represent the four combinations of preference and orientation comprising the cells of Table 1. A tabulation of the results illustrated in the graphs of Figures 1 to 3 is provided in Table 2. The layout of Table 2 allows for easy comparison with Table 1. Percentages are rounded to the nearest percent.

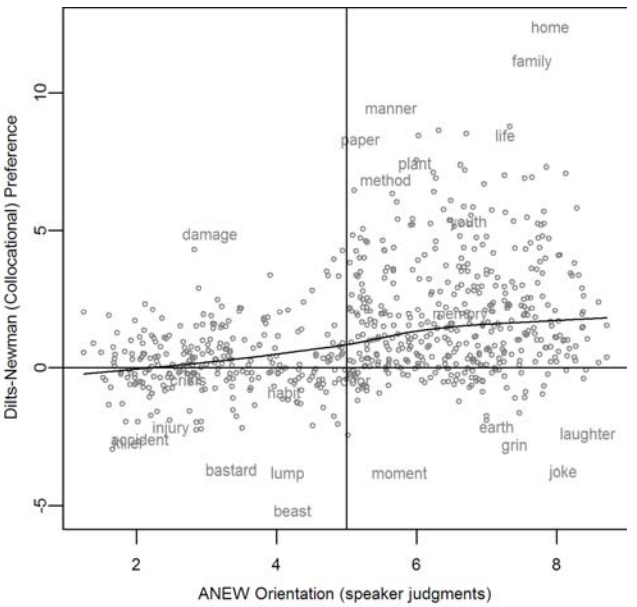


Figure 3. Semantic orientation vs. semantic preference for English nouns (full-strength seed adjectives)

Table 2. Number of nouns in each graph with each combination of semantic orientation and semantic preference.

Seed adjective set	Semantic preference	Semantic orientation	
		bad	good
empirical	positive	131 (18%)	355 (50%)
	negative	134 (19%)	97 (14%)
half-strength	positive	147 (20%)	379 (50%)
	negative	131 (17%)	79 (12%)
full-strength	positive	150 (20%)	370 (51%)
	negative	128 (18%)	88 (11%)

The most striking feature of Figures 1 to 3 is how much taller the top sections of the graphs are than the bottom sections—that is, how many more words have positive preference than negative preference—and how much stronger the positive preferences are. Table 2 shows that about 70% of the nouns in each analysis prefer to collocate with good adjectives, while around 30% prefer to collocate with bad adjectives.

The tendency is also shown by the curved lines running through the center of each graph. These are Lowess smoothers (Cleveland, 1981), or nonparametric regression measures. The lines appear to show modest correlations between semantic orientation and semantic preference, but these correlations are quite weak: Kendall's rank correlations (τ) are 0.217 for (1a), 2.59 for (1b), and 2.64 for (1c). If semantic orientation and semantic preference were more strongly correlated, we would expect these smoothers to cross to the top half of each graph at the vertical axes, representing neutral values. Instead, they all cross in the left-most quarter of their graph. That is, the average semantic preference becomes positive, while semantic orientations are still quite negative.

Indeed, in the full-strength and half-strength conditions, there are more bad nouns that prefer good adjectives than there are bad nouns that prefer bad adjectives. This stands in sharp contrast to the results of Dilts and Newman (2008), in which only one of the 13 bad nouns showed a preference for good adjectives, with the other 12 bad nouns preferring to collocate with reinforcing bad adjectives. While Dilts and Newman concluded that there was more transformation of the meaning of good nouns by bad adjectives, each of the present analyses contain more bad nouns being transformed by the meaning of good nouns: 18–20% of the nouns are bad nouns turning good, while only 11–14% of the nouns are good nouns turning bad. Why are the results of the present study the exact opposite of the results from Dilts and Newman in this respect? What has changed?

This difference in results is not due to a difference of opinion between the Dilts and Newman raters and the ANEW raters: 11 of the 50 nouns rated for "goodness" by Dilts and Newman's participants (plotted as words rather than dots in Figures 1 to 3) were also rated for "pleasantness" in ANEW, and the goodness and pleasantness ratings for all 11 of these nouns were consistent across the studies.

There is a more plausible explanation for the difference between the studies. Figures 1 to 3 show not only that there are more positive-preference nouns than negative-preference nouns but also that positive-preference nouns cover a wider range of preference scores than negative-preference nouns. That is, the points in the top halves of Figures 1 to 3 are more spread out vertically than the points in the bottom halves. More formally, positive-preference nouns have higher standard deviations than negative-preference nouns, as shown in Table 3 below.

In each Figures 1 to 3, it is not difficult to see that the words in the bottom-right quarter of the graph are clustered more tightly around the horizontal line than those in the bottom-left quarter of the graph, coinciding with the intuition that bad nouns should attract reinforcing bad adjectives, just as good nouns seem to prefer reinforcing good adjectives. The degrees of variation within the clusters of substantive data points in the bottom halves of the graphs are small, however, and are dwarfed by how far below those clusters the few outliers fall (*beast, laughter, bastard,* etc.). Looking only at the difference between these outliers would not reveal the difference between the more important, central data

points. By choosing the 25 most positive and negative nouns for analysis, Dilts and Newman started from the top and bottom of Figure 2 and proceeded toward the middle, selecting dots and effectively focusing their attention exclusively on these outliers.

Table 3. Standard deviations of the semantic preference of nouns in each graph with each combination of semantic orientation and semantic preference.

Seed adjective set	**Semantic preference**	**Semantic orientation**	
		bad	good
empirical	positive	0.63	1.19
	negative	0.55	0.65
half-strength	positive	0.76	1.60
	negative	0.68	0.67
Full-strength	positive	0.95	2.13
	negative	0.87	0.78

Looking further at the disparities in variances reveals an additional problem. Positive orientation nouns with positive preferences stand out in the top-right quadrant of Table 3. Their semantic preference shows more variance than any other category of noun. The variance of this section increases dramatically as the 62 adjectives with uncertain orientation are given more weight, as is apparent in Figures 1 to 3, where the points in the top-right quadrant dramatically increase in variance. While the standard deviation in other quadrants increases by at most 51% between the empirical and full-strength adjective conditions, the top-right quadrant shows a nearly 80% increase in deviation, to more than double that of all other quadrants.

A uniform increase in variance in all four quadrants might be a positive development, potentially allowing for finer discrimination in semantic preference among nouns. The disproportionate increase in variation for good orientation-positive preference nouns, on the other hand, suggests a systematic problem in the calculations getting worse as more adjective pairs are added.

Every possible noun in the BNC was considered for this study, suggesting that the increasing heteroscedasticity is not due to a lack of data, but to some deficiency in the method used to calculate collocational preference. In part, the difficulty may be due to the lower overall frequency of the negative adjectives in the adjective pairs that Osgood et al. (1957) chose to study.

The method for calculating preference in Dilts and Newman (2006) does take differences in frequency into account by using proportional frequencies. This technique breaks down when the proportional frequency of an adjective is zero, however. Consider the adjective pair *pure/impure*. *Pure* appears 3,301 times in the BNC, while *impure* appears only 68 times. By a simple ratio of proportional

frequencies means, if a word like *thoughts* is modified once by *pure* and zero times by *impure*, it will be considered completely positive, despite the fact that being modified only once by the highly frequent adjective *pure* seems in some way less likely than being modified zero times by an infrequent adjective like *impure*.

4. Conclusions and future work

The results illustrated in Figures 1 to 3 seem to suggest a correlation between semantic orientation and semantic preference: pleasant nouns tend to be preceded by pleasant adjectives, while unpleasant nouns tend to be preceded by unpleasant adjectives. In short, adjectives are chosen more to reinforce than to transform the nouns they modify. The analysis presented here does not thoroughly prove this tentative conclusion, but suggests that a study in which the range in semantic preferences considered is more carefully controlled might be able to do so.

All of the theoretical frameworks described in Section 2 assert, imply, or assume that semantic orientation and semantic preference should correlate. The computational studies described in Section 2.1.2 make the assumption that words with similar orientations tend to appear together, and the studies generally profit from this assumption. The semantic prosody studies described in Section 2.2 seem to focus on words with apparently neutral orientations that form constructions with positive or negative preferences, but no studies look for negative orientation words with positive preferences, or vice versa. If semantic prosody is more about diachronic process than it is about lexical properties, as suggested by Whitsitt (2005), then a word's semantic preference should pull its semantic orientation upward or downward over time, bringing these two factors into tighter correlation as time goes on.

A firmer conclusion could be reached by improving the method or materials used to calculate semantic preferences. First, new adjective pairs could be chosen from ANEW, for example, that are balanced for frequency as well as pleasantness (semantic orientation). Balancing frequency would eliminate the need to use proportional frequencies and would allow for continued use of the existing method for calculating collocational preference developed by Dilts and Newman (2006). Alternatively, a more sophisticated method for calculating collocational preferences could be developed or borrowed from the literature (e.g., Oakes, 1998).

5. Notes

* I would like to thank the members of the audience to which this paper was presented and two anonymous reviewers for their invaluable comments; John Newman, to whom this paper and I both owe a great deal; the editors of this volume for their suggestions and patience; and myself for any and

all mistakes. This work was supported by SSHRC Doctoral Fellowship number 752-2007-1311.

References

Baron, F., and G. Hirst (2003), "Collocations as cues to semantic orientation," unpublished m.s.

Bednarek, M. (2008), "Semantic preference and semantic prosody re-examined," *Corpus linguistics and linguistic theory,* 4: 119–139.

Bradley, M., and P. Lang (1999), *Affective Norms for English Words (ANEW): stimuli, instruction manual and affective ratings.* Technical report C-1, Gainesville, FL: The Center for Research in Psychophysiology, University of Florida.

Cleveland, W. S. (1981), "LOWESS: a program for smoothing scatterplots by robust locally weighted regression," *The American statistician*, 35: 54.

Dilts, P., and J. Newman (2006), "A note on quantifying 'good' and 'bad' prosodies," *Corpus linguistics and linguistic theory,* 2: 233–242.

Dilts, P., and J. Newman (2008), "Good nouns, bad nouns, and the company they keep," Presentation at the 2008 annual meeting of the LSA, Chicago, IL.

Hatzivassiloglou, V., and K. McKeown (1997), "Predicting the semantic orientation of adjectives," in: *Proceedings of the 35th annual meeting of the ACL and the 8th conference of the European chapter of the ACL.* New Brunswick, NJ: ACL, 174–181. Online at: <http://acl.ldc.upenn.edu/P/P97/P97-1023.pdf>.

Kendall, M. G. (1976), *Rank correlation method*. London: Charles and Griffin.

Louw, B. (1993), "Irony in the text or insincerity in the writer: the diagnostic potential of semantic prosody," in: M. Baker, G. Francis, and E. Tognini-Bonelli (eds.) *Text and technology: in honour of John Sinclair*. Amsterdam: John Benjamins, 157–174.

Oakes, M. P. (1998), *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Osgood, C., J. Suci, and P. Tannenbaum (1957), *The measurement of meaning.* Urbana: University of Illinois Press.

Partington, A. (2004), "Utterly content in each other's company," *International journal of corpus linguistics,* 9: 131–156.

Sinclair, J. (1987), *Looking up*: *an account of the COBUILD project in lexical computing*. London: Collins.

Sinclair, J. (1996), "The search for units of meaning," *Textus,* 9: 75–106.

Stefanowitsch, A., and St. Th. Gries (2003), "Collostructions," *International journal of corpus linguistics,* 8: 209–243.

Stone, P., D. Dunphy, M. Smith, D. Ogilvie, et al. (1966), *The General Inquirer: a computer approach to content analysis.* Cambridge: MIT Press.

Stubbs, M. (1995), "Collocations and semantic profiles: on the cause of the trouble with quantitative studies," *Functions of language,* 2: 23–55.

Stubbs, M. (2001), *Words and phrases: corpus studies of lexical semantics*. Oxford: Blackwell.

Turney, P., and M. Littman (2003), "Measuring praise and criticism: inference of semantic orientation from association," *ACM transactions on information systems,* 21: 315–346.

Whitsitt, S. (2005), "A critique of the concept of semantic prosody," *International journal of corpus linguistics,* 10: 283–305.

Wiebe, J., T. Wilson, and C. Cardie (2005), "Annotating expressions of opinions and emotions in language," *Language resources and evaluation,* 39: 165–210.

Wilson, T., J. Wiebe, and P. Hoffmann (2005), "Recognizing contextual polarity in phrase-level sentiment analysis" in: *Proceedings of the human language technology conference and conference on empirical methods in natural language processing,* 347–354. Online at: <http://www.cs.pitt.edu/~wiebe/pubs/papers/emnlp05polarity.pdf>.

Wurm, L., and D. Vakoch (2000), "The adaptive value of lexical connotation in speech perception," *Cognition and emotion,* 14: 177–191.

Wurm, L., D. Vakoch, S. Seaman, and L. Buchanan (2004), "Semantic effects in auditory word recognition," *Mental lexicon working papers,* 1: 47–62.

Wurm, L. (2007), "Danger and usefulness: an alternative framework for understanding rapid evaluation effects in perception?" *Psychonomic bulletin and review,* 14: 1218–1225.

Xiao, R., and T. McEnery (2006), "Collocation, semantic prosody, and near synonymy: a cross-linguistic perspective," *Applied linguistics,* 27: 103–129.

Subject omission in Russian: a study of the Russian National Corpus

Tatiana Zdorenko

University of Alberta

Abstract

This study investigated subject omission in spoken and written corpora of Russian in order to produce a quantitative comparison of omission in different genres and morphosyntactic environments. Previous theoretical studies of Russian described subject omission using isolated constructed sentences, and most corpus studies analyzed written literary language. Since subject omission in Russian is a discourse phenomenon, the present study investigated subject omission in coherent spontaneous text, focusing on spoken data from the Russian National Corpus. In the corpus, subjects were not omitted to the same extent in all genres and registers. The percentage of omitted subjects was the highest in the corpus of informal spontaneous conversations, and omitted subjects were practically absent in the written corpus, even in the most informal register. The comparison of the frequency of null subjects in different person contexts provided support for the Topicalization Hierarchy of person. More subject omission was found in the first- and second-person contexts than in the third-person contexts. In contrast, in written Russian there was no significant effect of person on the proportion of null subjects. Finally, an analysis of omitted subjects used with specific verb types was built on previous cross-linguistic studies of grammaticalized collocations such as I dunno *or* y'know. *It was concluded that* znat' *'know' and* ponimat' *'understand' were likely candidates for grammaticalization as discourse markers, i.e., verbs with a particular pragmatic function that grammaticalized in a subjectless form.*

1. Introduction

The phenomenon of subject omission is one of the most studied topics in theoretical syntax. However, in previous studies, theorists have mostly focused on creating an elegant cross-linguistic typology of subject-omission phenomena using constructed examples rather than spoken language corpora. Less attention has been given to quantitative properties of subject omission within one language and across languages. There are still very few studies that use large representative samples of spontaneous speech to investigate subject omission.

Rizzi (1982) was one of the first to conduct a study that analyzed subject omission in the light of the principles and parameters framework, and his study triggered an extensive investigation of the phenomenon in other languages within the generative framework. In his study of Italian syntax, Rizzi hypothesized that null subjects in Italian can be accounted for in terms of the pro-drop parameter. He proposed that the positive or negative setting of the pro-drop parameter could describe the distinction between languages that allow subject omission and languages that disallow it. Initially, it was assumed that all languages of the world

could be classified into two groups, pronoun-drop (pro-drop) languages and non-pro-drop languages. For instance, Italian, Spanish, and Greek were classified as pro-drop languages, while English and French were considered non-pro-drop languages. This cross-linguistic distinction was first explained in terms of the richness of the verb inflection. The main idea of this approach can be described as follows: if in a language every person and number form in the paradigm is uniquely identified by a suffix on the verb, syntactic subjects in such a language need not be overt. This idea runs into problems if we look at such Asian languages as, for instance, Mandarin and Cantonese (Li and Thompson, 1981; Matthews and Yip, 1994). These languages have no subject-verb agreement morphology but nevertheless allow subjects to be dropped. In his studies of Mandarin Chinese, Huang (1984, 1989) proposed that subjects in this language and other similar languages can be dropped when they are uniquely identified by an antecedent noun phrase in discourse or by a salient referent in the situational context. In other words, omitted subjects in such languages are omitted topics.

Then it becomes clear that verb inflection and the pro-drop parameter are not sufficient to capture the distribution of subject types across languages, for at least two main reasons: (1) it is not always possible to account for subject omission in pro-drop languages based on verb inflection alone, and (2) in all languages, morphologically rich or not, subject omission cannot be viewed separately from information structure in discourse. For example, English, a non-pro-drop language, permits null subjects in diaries or informal written style (Haegeman, 1990; Haegeman and Ihsane, 1999, 2001). However, many studies still build their analyses of subject omission on the dual distinction between pro-drop and non-pro-drop languages. This is the case in the studies of the acquisition of Russian (Bar-Shalom and Snyder, 1997; Gordishevsky and Avrutin, 2003, 2004) as well as the studies of adult language (Franks, 1995; Seo, 2001).

Russian is particularly interesting with respect to the phenomenon of subject omission, because it is a topic-prominent language in which the information structure plays an important role in the organization of an utterance. This property is demonstrated in flexible word order and the possibility of omitting elements that are semantically redundant. Discourse omission is illustrated in the following dialogue taken from the Russian National Corpus:

(1) *My* *vstret-im-s'a?*
we meet-pres.1pl-refl?
'Shall we meet?'
Davajte, *kak* *i* *dogovariv-al-i-s'.*
let.us as and agree-past-pl-refl
'Let's do it, as (we) already agreed'

Note that in the second line of the dialogue, the verb form does not mark person agreement, so the subject is not identified by the verb suffixes. The subject *my* 'we' was omitted because the preceding question identifies the subject. However, even though in this particular case the subject was omitted, it is

possible to use an overt subject in the answer: *kak **my** i dogovarivalis'* 'as **we** already agreed.' Such an utterance would not be redundant or emphatic. Most studies of null subjects in Russian seem to have come to a conclusion that subject omission in this language is a manifestation of optional topic-drop.

2. Previous studies of subject omission in Russian

2.1 Theory-based studies

Russian presents an interesting case for a study of subject omission. On the one hand, subject pronouns can be overt without being emphatic. This is not the case in languages like Italian or Spanish, in which overt pronouns are used only for emphasis or disambiguation. Despite rich verb agreement, which in most cases can identify the referent of the subject, overt subjects in Russian are used more often than in Romance null-subject languages. According to estimations reported in earlier studies, the subject omission rate in Russian is approximately 20% (Gordishevsky and Avrutin, 2004; Grenoble, 2001). In contrast, in "canonical" null-subject languages like Spanish or Italian, 70–80% of subjects are omitted (Davidson, 1996; Valian, 1991).

A formal syntactic study of subject omission across Slavic languages (Franks, 1995) comes to a similar conclusion. According to Franks, East Slavic languages, Russian and Ukrainian, have more overt subjects than other Slavic languages. Thus, he argues that while subject omission in languages like Polish is a case of pro-drop licensed by verb agreement, subject omission in Russian could be the result of "surface deletion," or in other words, omission of the topical constituent in an utterance (292). However, it is not possible to investigate this issue further using examples given in Franks because his work has the limitations of a theory-driven study, one of them being that it uses constructed examples. Since subject omission is in most cases discourse-based, single sentences taken out of context are not sufficient to evaluate the acceptability of a null argument.

In the discussion of the contexts of subject omission in the speech of Russian children, Gordishevsky and Avrutin (2003, 2004) argued that subjects can be omitted provided that certain discourse conditions are satisfied. The two main conditions for omission are the presence of a linguistic antecedent in one of the previous utterances or the presence of a salient situational antecedent. They also make a distinction between the first and second person and the third person, because first- and second-person subjects always have a situational antecedent, i.e., the speaker or the hearer. An advantage of the data in this study compared to Franks (1995) is that the authors constructed mini-dialogues for each type of omission context, which is a better way of describing a discourse-bound phenomenon. In (2) and (3) below, I quote examples from Gordishevsky and Avrutin (2003: 5–6) illustrating two contexts of omission:

(2) Linguistic antecedent (3rd person subject)
a. *Gde* *Ivan?*
where Ivan
'Where is Ivan?'
b. *Ush'o-l* *domoj.*
went-sg:past home
'(He) went home'

(3) Situational antecedent (1st person subject)
Hoch-u *jabloko.*
want-1sg:pres apple
'(I) want an apple'

The subject is omitted from the response in (2b) because its referent is present in the question. In (3), the subject, i.e., the speaker, is present in the situational context. These data, although grammatically correct, still consist of constructed examples and thus do not represent actual spontaneous usage. Such examples are perhaps sufficient to illustrate the idea that subject omission in Russian depends on the contextual status of the subject, i.e., whether the subject has a salient antecedent in the preceding discourse or in the situational context. However, in order to study a phenomenon such as subject omission, a more appropriate approach is to analyze samples of coherent spontaneous text, and this was the approach taken in corpus studies of subject omission and in the present study.

2.2 Corpus-based studies of Russian

Subject omission in Russian has not been widely investigated using corpora. Theorists turned to data from spoken corpora only recently, perhaps because earlier studies were influenced by the idea of written literary language as the norm. For instance, Kibrik (1996) analyzed the distribution of subject types in one short story. Seo (2001) conducted a study that used a much wider selection of literary sources in the analysis of subject omission. The author based his analysis on novels either originally written in one of five Slavic languages (Russian, Polish, Czech, Bulgarian, and Serbo-Croatian) or translated into these languages. A sample of 2,000 clauses was selected from the novels for each of the five languages, resulting in a corpus of 10,000 clauses in total. Interestingly, in the selection of the sample, the author included those parts of the novels that contained most dialogue. Seo then compared proportions of null subjects in various morphosyntactic environments across the five languages and found 22% omitted subjects in Russian compared to 80–90% omitted subjects in other Slavic languages, which supports the claim in Franks (1995) that Russian is different from other Slavic languages. However, Seo's (2001) conclusions should be interpreted with caution, not only because the study was based on a written corpus, but also because parts of the corpus were translations into Russian from other Slavic languages.

To my knowledge, Grenoble's (2001) study is the only study of null subjects in spoken Russian. In this study, 45 minutes of taped conversations were analyzed with regard to the use of nominal and pronominal subjects and subject omission. Comparing the findings of the study of a written corpus (Seo, 2001) and a spoken corpus (Grenoble, 2001), we can see that the overall proportion of omitted subjects reported in the two studies is very similar (approximately 22%). However, the findings of the two studies differ with regard to the proportions of null subjects in different contexts. Grenoble found an effect of person on subject omission (19% in first-person vs. 43% in second-person and 17% in third-person contexts). Seo (2001) found no effect of person, as the proportion of null subjects was 21–23% in first-, second-, and third-person contexts in Seo's written corpus. Grenoble (2001) also makes an interesting note that such verb forms as *znaesh* '(you) know' and *ponimaesh* 'you understand' can be fixed expressions that are always used with a null subject. I will return to the issue of fixed expressions later in this paper.

In sum, the two corpus studies of Russian found that, in addition to discourse factors, the choice of subject form also depends on morphosyntactic factors, but the evidence regarding the role of each of these factors is contradictory. Even though the two studies compared in this section found similar overall proportions of null subjects, the sources of data and inclusion criteria were very different. For these reasons, and also because the two studies analyzed very different genres of colloquial and literary language, one should be cautious comparing their results and generalizing their conclusions.

3. A study of subject omission in the Russian National Corpus

3.1 Research questions

Previous studies mostly used written sources or constructed examples to make conclusions about the extent of subject omission in Russian, but a better way to describe this phenomenon would be to compare the proportion of subject omission in written and spoken language. The research questions to be answered in this study are the following:

1. Is the rate of subject omission higher in spoken genres than in written genres? Previous corpus studies did not find a difference between the two, but since subject omission in Russian appears to be the result of ellipsis, I expect to find more omitted subjects in spoken language.
2. Is the rate of subject omission higher in informal genres, written and spoken? Again, if subject omission is a discourse phenomenon, I expect to find more omitted subjects in blogs compared to fiction, or in conversations compared to lectures.

3. What are the rates of subject omission in different person contexts? In line with the previous theoretical studies, I expect that subjects are omitted more often in the first- and second-person contexts.

3.2 The Russian National Corpus (RNC)

The data I used to answer the research questions were drawn from the Russian National Corpus (<http://ruscorpora.ru>), a project designed by the Institute of the Russian Language at the Russian Academy of Sciences. The RNC is a representative collection of Russian texts in electronic form that are publicly available from the web site. The present version of the corpus includes over 140 million words. The RNC is representative of various genres and registers and includes two major parts, a written corpus and a spoken corpus. The written corpus in turn consists of several subcorpora, such as fiction, newspaper texts, academic writing, personal correspondence, etc. The spoken corpus includes transcripts of conversations that are subdivided into two subcorpora: public speech, such as lectures and interviews, and private conversations, such as informal dialogue and storytelling.

Recall that the previous studies did not compare subject omission in different genres but rather chose one genre, spoken or written, usually focusing on the written genre. Taking into account this limitation, I decided to carry out analyses of subject omission in both spoken and written Russian and, furthermore, compare different registers within spoken and written language. I selected three subcorpora from the written corpus:

1. fiction (detective novels published in the period from 1990 to 2006);
2. newspaper texts (news reports and commentaries, mostly from the late 1990s and early 2000s);
3. electronic communication (blogs, early 2000s).

I also selected the following three subcorpora from the spoken corpus:

- lectures from several Russian universities (1990-2005);
- interviews and press-conferences (early 2000s);
- informal private conversations (transcripts of personal and telephone dialogs recorded in 1990–2005).

The idea behind the selection of particular subcorpora from the spoken and the written corpus was to cover various degrees of formality in both genres. The three subcorpora in each genre were chosen to represent a range from the most formal and premeditated register, such as lectures in the case of the spoken subcorpus, to the most informal register, such as Internet blogs in the case of the written subcorpus.

3.3 Coding procedure

Identifying sentences with omitted subjects was not an easy task for several reasons. Since the corpus was not marked up for syntactic relations, it was impossible to search for subject positions in clauses automatically. Moreover, even though the neutral word order in Russian is Subject Verb Object, the order of arguments is flexible and the beginning of the sentence is not necessarily the location of the subject as it is in English. In order to somewhat facilitate the analysis, I searched the selected subcorpora for all utterances containing a verb. The search was further limited to verbs in the indicative mood in order to exclude imperative utterances, which represent a special kind of subjectless utterance. I excluded impersonal constructions that require a null subject, such as *mne kazhetsja* '(it) seems to me' or *govorjat* '(they) say' from the analysis.

The search returned thousands of hits for each subcorpus, but only the first 100 hits (i.e., sentences with a verb in the indicative mood) were coded for the initial analysis. These 100 sentences represented an unbiased sample of the search results for the following reasons: texts that contained the search hits were not sorted by title, the year of publication or recording, or the location and represented spoken and written Russian from the 1970s to the 2000s. I selected the first 10 hits from each of these texts, in order to make sure that in each genre, 100 sentences originated from several texts rather than one long text. In the case of some conversations, blogs, and newspaper articles, there were fewer than 10 hits and all the utterances were included. An anonymous reviewer pointed out that, since subject omission is a discourse phenomenon, it can be correlated with discourse organization. It is indeed very likely that subject-omission rates can change depending on what part of a story or conversation we analyze, but in the present paper, I aimed to investigate overall rates of subject omission. Furthermore, since most of the texts from the informal genres were very short and returned fewer than 10 hits, they were included in the analysis in their entirety. Thus, I did not limit my analysis to utterances from the initial parts of texts.

The utterances that met the criteria for inclusion were coded for the subject type (nominal, pronominal, or zero), person and number of the subject, tense, and clause type (main or subordinate, declarative or interrogative). For the purposes of this paper, I will focus only on the subject type and person analysis.

4. Results

4.1 Comparative analysis of genres in the spoken and written RNC

The aim of the analysis of subject omission in six subcorpora from the spoken and the written RNC was to answer the following research questions: (1) Is the rate of subject omission related to genre, and if so, is it higher in the spoken corpus? (2) Within one genre, is the subject omission higher in more informal registers, which are possibly more elliptical? In Tables 1 and 2, I summarize the tokens and percentages of subject types in the spoken and the written subcorpus respectively. The columns in each table represent three different registers in the subcorpus.

Table 1. Proportions (tokens) of subject types in three subcorpora of the spoken RNC.

Subject Type	**Genre**		
	private	interview	lecture
noun	35% (28)	59% (42)	58% (42)
pronoun	33% (26)	35% (25)	39% (28)
null	32% (25)	6% (4)	3% (2)

Table 2. Proportions (tokens) of subject types in three subcorpora of the written RNC.

Subject type	**Genre**		
	blogs	newspapers	fiction
noun	39% (23)	82% (60)	74% (53)
pronoun	59% (35)	14% (10)	26% (19)
null	2% (1)	4% (3)	0% (0)

Even a simple comparison of the proportions of null subjects in different genres (the last row in the tables) indicates that, first of all, null subjects occurred more often in spoken genre than in written genre. In spoken genre, 31 out of 222 subjects were null, while in written genre, only 4 out of the total of 204 subjects were null. Moreover, 3 instances out of these 4 occurred in interviews that were quoted in the reportages. In other words, out of 35 observed null subjects, 31 occurred in speaking and only 4 in writing. The results of chi-square tests revealed significant distributions in Table 1 ($\chi^2 = 33.53$, $df = 4$, $p < .001$) and Table 2 ($\chi^2 = 35.62$, $df = 4$, $p < .0001$), which are due to the high number of null subjects in private conversations (Table 1) and the high and low numbers of pronouns in blogs and newspapers respectively (Table 2).[1]

The second observation concerns the distribution of subject types across subcorpora *within* spoken genre (Table 1). Note that the null subjects were not distributed evenly across the three subcorpora: most of them occurred in private

conversations (25 null subjects out of the total of 31). This leads us to the important conclusion that subject omission appears to be a feature of spoken language and, moreover, a feature of *informal* spoken language. Thus, coming back to the research questions we asked at the beginning of this section, we can say that speech genre and register certainly influence the extent of subject omission and that it is present in spoken language but not in written language. The findings also support our prediction that in informal and less scripted situations the speaker relies more on shared knowledge and on the common situational context of the speaker and the hearer, which may lead to more ellipsis in speech.

4.2 Subject omission in informal conversation corpus

Since a considerable proportion of null subjects was found only in the sample of informal conversations, I decided to further limit investigation of subject omission to this particular genre of the spoken RNC. Following the coding procedure described in Section 3.3, I selected a larger sample of 600 utterances from the conversational subcorpus for a more detailed analysis. Table 3 summarizes the percentages and tokens of subject types in main clauses in my sample of the RNC and also compares the numbers with those reported by Grenoble (2001: 9) and by Seo (2001: 165–67). In order to make a comparison of the three studies possible, I report only the numbers of pronominal and null subjects and exclude nouns, following Seo's (2001) practice.

Table 3. Percentages (tokens) of pronominal and nominal subjects the Russian National Corpus and in earlier corpus studies.

Subject type	**Study**		
	RNC	Grenoble (2001)	Seo (2001)
pronoun	71% (254)	76% (351)	88% (1,557)
null	29% (101)	24% (111)	22% (443)

Chi-square tests confirmed that there is a significant difference ($\chi^2 = 8.18$, $df = 1$, $p < 0.01$) between the number of pronominal and omitted subjects in the RNC and in Seo (2001). There was no significant difference between the numbers of pronominal and omitted subjects in the present study and those in Grenoble's (2001) study ($\chi^2 = 3.81$, $df = 1$, $p > 0.05$).[2] Recall that Grenoble's study was of spontaneous conversations, and Seo (2001) used novels as the source of data. Even though Seo included samples of novels that contained more dialogue, the overall proportion of omitted subjects in the written texts still did not reach the proportion characteristic of informal spoken register. This is not surprising because dialogue in fiction by no means represent spontaneous speech. This again leads us to the conclusion that certain features of written language should not be generalized to the language in general.

4.3 Person, topic, and subject omission

The role of person and animacy in the information structure was summarized by Givón (1976) in the Topicalization Hierarchy:

(4) 1st person, 2nd person > 3rd person
human > animate > inanimate

The Topicalization Hierarchy reflects higher topicality of human arguments in general and the first and second person in particular, and it is manifested across languages in several syntactic and discourse phenomena, such as word order and grammatical relations, as well as argument omission.

Having established the overall percentage of null subjects in the spoken corpus in the previous section, the next task in the analysis of null subjects was to investigate the role of person in subject omission. The relation between person and topicality lead us to the following research question: In spoken language, how are null subjects distributed by person? I expected person to have an effect on the proportion of subject omission, because different person forms have different status with respect to the distribution of old and new information in the discourse. Third-person subjects can denote both old and new information and can be accordingly represented as pronominal subjects, null subjects, or full noun phrases. First- and second-person subjects, on the other hand, are always topics, i.e., they always have a discourse antecedent, because the identities of the speaker and the hearer are taken as a background assumption in any conversation. In other words, are there differences in the proportions of subject omission in various person contexts? If person and the Topicalization Hierarchy play a role in subject omission in Russian, we expect more null subjects in the first and second person and fewer in the third person. The distribution of subject types in different person contexts in the RNC is summarized in Table 4.

Table 4. Types of subjects used in first-, second-, and third-person contexts in the RNC.

Subject type	**Person**		
	1	2	3
null	27% (46)	33% (23)	12% (32)
pronoun	73% (126)	67% (46)	32% (82)
noun	-	-	56% (147)

Comparing the proportions of null subjects in the spoken RNC and in written Russian, I again found important differences. Seo (2001) established that in the corpus of novels, the percentages of null first-, second-, and third-person subjects were the same (22%). The data from the RNC are thus more compatible with the findings of Grenoble's (2001) study of spoken Russian, which found more null subjects in the second-person (43%) than in the first- and third-person contexts (19% and 17% respectively) in conversational Russian. Coming back to the research questions asked in this section, it can be concluded that the data from

conversational Russian provided support for the Topicalization Hierarchy. Indeed, person seemed to play a role in the proportion of subject omission, and the percentage of null subjects was higher in the first and second person than in the third person.

4.4 Discourse markers

One of the topics that remain undiscussed in previous studies of subject omission in Russian is the analysis of particular verb types used with null subjects, as well as the relation between grammaticalization of verb forms and subject omission. In some grammaticalized constructions, verbs tend to be tied to a specific subject form, i.e., tend to be used exclusively with an overt subject or with a null subject. As a result, subject usage can be an integral part of such fixed expressions, where the speaker has no choice of a null or an overt subject but rather uses the construction as an unanalyzed chunk. Such fixed expressions are known in the literature as "epistemic markers" (Thompson and Mulac, 1991) or "discourse markers" (Fraser, 1990 and Schiffrin, 1987, among others). Consider the following examples of discourse markers (DMs): the Spanish *no sé* '(I) don't know' (Davidson, 1996), the Estonian *tea* '(you) know' or *ei tea* '(you) don't know' (Keevalik, 2006), and *remember* in English (Tao, 2003).

Discourse markers (DMs) are independent elements that are used to structure a conversation and maintain cohesiveness. For instance, a DM can be used to initiate discourse, change the topic, hold or change turns, express the speaker's attitude to an utterance, or indicate the relationship between an utterance and preceding discourse. DMs can be independent words or can evolve from larger phrases or entire constructions. DMs are therefore often considered to be an instance of grammaticalization whereby the semantic content of the original expression becomes bleached and the phrase acquires a distinct pragmatic meaning sometimes not related to the original meaning of the elements of the DM. Importantly, when a verb constitutes a part of a DM, it may become grammaticalized in a null-subject form, which is the case in all the examples in the previous paragraph. DMs often become phonologically reduced (e.g., *I dunno* or *y'know* in English) and syntactically flexible, i.e., they can occur at any point in a syntactic structure.

In order to isolate the potential discourse markers in the spoken RNC, I carried out an analysis of the distribution of verb lemmas that occurred with overt and omitted subjects. The aim of this analysis was to find specific verbs that frequently occur in a particular person form with a null subject or with a pronominal subject. Verbs that occurred only in second-person forms in the sample were *znat'* 'know' and *ponimat'* 'understand.' Their forms *znaesh* '(you) know' and *ponimaesh* '(you) understand' were used as independent elements sentence-initially. They were often preceded and followed by a pause (marked as a "/" in the corpus), again indicating that these expressions are independent elements and can be taken out without changing the meaning of the utterance. In (5) and (6), I provide examples of used as discourse markers:

(5) *No / ponimaesh / u menja nemetskij osnovnoj.*
But understand-2sg at/on me German major
'But, (you) see, German is my major'

(6) *Nu / znaesh / tam iz sotni fotok neskol'ko udachnyh.*
well know-2sg there from hundred photos several successful
'Well, (you) know, there is a couple of good photos out of a hundred'

These uses of the verbs are different from the one given below. Note that in (5) and (6), *znaesh* '(you) know' and *ponimaesh* '(you) understand' are used without a subject. In contrast, *ponimaesh* '(you) understand' as a main verb in (7) has a subject and an object.

(7) *Ty ne ochen' menja ponimaesh.*
You not very me understand-2sg
'You don't quite understand me'

Thus, *znat'* 'know' and *ponimat'* 'understand' indeed possess the properties of a DM when they are used without a subject. However, these two verbs were not identical in their patterns of use. The verb 'understand' mostly occurred with a null subject in second-person form, as in (5). The verb *znaesh* '(you) know' occurred with a null subject, as in (6), but also with the pronominal subject 'you,' with and without an object. However, my sample of 600 utterances from the conversational subcorpus of the RNC was not sufficient to compare only these two verbs. I searched the entire conversational subcorpus for all instances of *znaesh* and *ponimaesh*. The results are summarized in Table 5.

Table 5. Percentages (tokens) of subject types used with *znaesh* '(you) know' and *ponimaesh* '(you) understand' in the RNC.

Subject type	**Verb**	
	znaesh	*ponimaesh*
null	49% (145)	98% (125)
pronoun	50% (147)	2% (2)

The results of the search in the entire conversational subcorpus confirmed the tendency found in the sample. The form *ponimaesh* appeared to be more grammaticalized as a DM than *znaesh*, since *ponimaesh* was used with a null subject in almost 100% of cases (and most of the 125 instances were DM uses), while *znaesh* was used without a subject in only 50% of cases. The results of a chi-square test confirmed that the distribution summarized in Table 5 was highly significant ($\chi^2 = 89.74$, $df = 1$, $p < .0001$).

Another sign of grammaticalization of *ponimat'* 'to understand' in the second-person form *ponimaesh* as a DM is the fact that this verb form has acquired a slightly different shade of meaning, similar to the English verb *see* used in the expression *(you) see*. Most uses of *ponimaesh* '(you) understand/

(you) see' are clearly pragmatic and are aimed at simply engaging the hearer in the conversation rather than literally making sure that the utterance is understood.

In sum, I found that among the verbs frequently used in discourse, *ponimat'* 'understand' is, in turn, more often used as a DM than *znat'* 'know,' and as a result, it appears more often in a subjectless form than *znat'*. Tying this finding to the finding regarding the distribution of null subjects in first-, second-, and third-person contexts, I can conclude that the presence of such DM in conversation could be another factor contributing to the high rate of null subjects in second-person contexts.

5. Conclusion

This study has investigated subject omission in spoken and written corpora of Russian with the goal of making a quantitative comparison of omission in different genres and morphosyntactic environments. An investigation of small samples of various genres represented in the RNC lead me to the conclusion that subjects were not omitted to the same extent in all genres and registers. I found that the percentage of null subjects was the highest in the corpus of informal spontaneous conversations and that null subjects were practically absent in the written corpus, even in the most informal register. This finding again provides evidence against the simplistic division of all languages into null-subject and non-null-subject languages. A new fact that has been overlooked in previous work on Russian is that subject omission in this language is in fact only a feature of a particular speech register, namely informal unscripted conversation.

In a more detailed analysis of the spoken corpus sample, I examined the relation between null subjects and the notion of subject topicality. The comparison of the frequency of null subjects in different person contexts provided support for the Topicalization Hierarchy of person. More null subjects were found in the first- and second-person contexts than in the third-person contexts, in contradiction with a study of written Russian, which found no significant effect of person on the proportion of null subjects.

Finally, an analysis of null subjects used with certain verb types built on the previous studies of grammaticalized collocations such as *I dunno* or *y'know*. The analysis was aimed at isolating potential discourse markers, i.e., verbs with a particular pragmatic function that grammaticalized either in a subjectless form or with a certain pronominal subject. I found that *znat'* 'to know' and *ponimat'* 'to understand' were the verbs whose second-person forms often functioned as discourse markers. However, further research and larger speech samples are needed to detect other discourse markers in Russian (e.g., first-person forms of the same verbs) and to investigate their pragmatic functions in more detail.

6. Notes

1. I would like to thank the anonymous reviewer for carrying out the statistical evaluation of the data in Table 1 and Table 2.
2. I am grateful to the anonymous reviewer for pointing out the error in my original computation of the chi-square test.

References

Bar-Shalom, E., and W. Snyder (1997), "Optional infinitives in child Russian and their implications for the pro-drop debate," in: M. Lindseth and S. Franks (eds.) *Formal approaches to Slavic linguistics: the Indiana meeting 1996*. Ann Arbor: Michigan Slavic Publications.

Davidson, B. (1996), "Pragmatic weight and Spanish subject pronouns: the pragmatic and discourse uses of 'tú' and 'yo' in spoken Madrid Spanish," *Journal of pragmatics,* 26: 543–565.

Franks, S. (1995), *Parameters of Slavic morphosyntax*. Oxford: Oxford University Press.

Fraser, B. (1990), "An approach to discourse markers," *Journal of pragmatics,* 14: 383–395.

Givón, T. (1976), "Topic, pronoun, and agreement," in: C. N. Li (ed.) *Subject and topic.* New York: Academic Press, 149–188.

Gordishevsky, G., and S. Avrutin (2003), "Subject and object omissions in child Russian," in: Y. N. Falk (ed.) *Proceedings of IATL 19*. Online at: <http://linguistics.huji.ac.il/IATL/19/GordishevskyAvrutin.pdf>.

Gordishevsky, G., and S. Avrutin (2004), "Optional omissions in an optionally null subject language," in: J. van Kampen and S. Baauw (eds.) *Proceedings of GALA 2003: Volume 1*. Groningen, The Netherlands: LOT, 187–198.

Grenoble, L. (2001), "Conceptual reference points, pronouns, and conversational structure in Russian," *Glossos,* 1(1). Online at: <http://www.seelrc.org/glossos/issues/1/grenoble.pdf>.

Haegeman, L. (1990), "Non-overt subjects in diary contexts," in: J. Mascaro and M. Nespoor (ed.) *Grammar in progress*. Dordrecht: Foris, 167-174.

Haegeman, L., and T. Ihsane (1999), "Subject ellipsis in embedded clauses in English," *English language and linguistics,* 3(1): 117–145.

Haegeman, L., and T. Ihsane (2001), "Adult null subjects in the non-pro-drop languages: two diary dialects," *Language acquisition,* 9: 329–346.

Huang, C.-T. J. (1984), "On the distribution and reference of empty pronouns," *Linguistic inquiry,* 15: 321–337.

Huang, C.-T. J. (1989), "Pro-drop in Chinese: a generalized control theory," in: Osvaldo Jaeggli and Ken Safir (eds.) *The null subject parameter*. Boston: Kluwer, 185-214.

Keevalik, L. (2006), "From discourse pattern to epistemic marker: Estonian *(ei) tea* 'don't know,'" *Nordic journal of linguistics,* 29: 173–200.

Kibrik, A. (1996). "Anaphora in Russian narrative prose: a cognitive calculative account," in: B. Fox (ed.) *Studies in anaphora.* Amsterdam: John Benjamins. 266–303.

Li, C. N., and S. Thompson (1981), *Mandarin Chinese: a functional reference grammar*. Berkeley, CA: University of California Press.

Matthews, S., and V. Yip (1994), *Cantonese: a comprehensive grammar*. London: Routledge.

Rizzi, L. (1982), *Issues in Italian syntax*. Dordrecht: Foris.

Schiffrin, D. (1987), *Discourse markers*. Cambridge: Cambridge University Press.

Seo, S. (2001), "The frequency of null subject in Russian, Polish, Czech, Bulgarian, and Serbo-Croatian: an analysis according to morphosyntactic environments," Ph.D. dissertation, Indiana University at Bloomington.

Tao, H. (2003), "A usage-based approach to argument structure: 'remember' and 'forget' in spoken English," *International journal of corpus linguistics,* 8(1): 75–95.

Thompson, S. A., and A. Mulac (1991), "A quantitative perspective on the grammaticalization of epistemic parentheticals in English," in: E. Traugott and B. Heine (eds.) *Approaches to grammaticalization*. Amsterdam: John Benjamins, 313-339.

Valian, V. (1991), "Syntactic subjects in the early speech of American and Italian children," *Cognition,* 40: 21–81.

Linguistic realizations of rhetorical structure: a corpus-based study of research article abstracts and introductions in applied linguistics and educational technology

*Phuong Dzung Pho**

Monash University, Melbourne, Australia

Abstract

The abstract and introduction of an article are at the forefront of an article. They are the first parts of an article to be read by the reader. How to write good abstracts and introductions can be challenging to novice writers. Although there has been considerable research on the rhetorical structure of research articles, these studies tend to focus on the generic structure or move structure of the article. What is equally important, and perhaps more useful to novice writers, is how each move can be realized linguistically. Most of the previous studies in this area are limited in that they focus on the distribution patterns of only one or two linguistic features in either the abstract or the introduction of the main article as a whole rather than examine the distribution of a more comprehensive set of features at the move level. Furthermore, it is sometimes hard for novice writers to distinguish the way they should write the abstract, which precedes the article and is independent of the article, and the introduction of the main article. These two genres are seemingly similar, yet they have distinctive move structures and linguistic characteristics. This study thus aims at examining a range of linguistic features of each of the abstract and introduction moves of research articles in two disciplines, applied linguistics and educational technology. A corpus of 40 research articles in these two disciplines was xml-tagged for moves and coded for a range of linguistic features to investigate what features are prototypical of each move. The analysis shows that a combination of features such as verb tenses, voice, modal verbs, stance words, self-reference words, and reporting verbs can help to distinguish moves. Variations across the two disciplines are also examined. These findings have pedagogical implications for academic writing courses for graduate students in general and for students from non-English backgrounds in particular.

1. Introduction

That nonnative speakers of English experience difficulty in having their papers published in international English-medium journals has been well documented in the literature (Burrough-Boenisch, 2003; Cho, 2004; Flowerdew, 1999; Gosden, 1992b; Lillis and Curry, 2006; Misak, Marusic, and Marusic, 2005; Pagel, Kendall, and Gibbs, 2002; Sionis, 1995). There has been a considerable number of studies of academic and professional writing, research articles in particular. Most of these studies, however, seem to focus on the rhetorical structure or move structure of the research article rather than the linguistic realizations of the rhetorical moves (e.g., Posteguillo, 1999; Samraj, 2002; Yang and Allison, 2004). *Moves* refer to segments of text with a certain function (Swales, 1990); for

example, a segment of text in the introduction section of a paper that establishes a niche in the field is called an *Establishing a niche* move. Although knowing the move structure of an article is useful to novice writers, knowing the conventional linguistic realizations of various rhetorical functions, as pointed out by Graetz (1985) and Ventola (1994), is no less important. Some studies also investigate the linguistic features of research articles, but usually focus on only one or two features, such as voice (Martínez, 2001; Tarone, Dwyer, Gillette, and Icke, 1998), tenses (Burrough-Boenisch, 2002; Malcolm, 1987), reporting verbs (Hyland, 2002; Thompson and Ye, 1991), evaluation and hedging (Hood, 2005; Hyland, 1996; Tucker, 2003; Vassileva, 2001), personal pronouns (Harwood, 2005; Kuo, 1999), and theme (Gosden, 1992a; Martínez, 2003).

Most studies of the linguistic features of the research article genre observe the distribution patterns of the features across sections rather than moves. For example, Martínez (2001) observes the distribution of active and passive voice in the Introduction, Methods, Results, and Discussion sections. Although it may be true that linguistic realizations vary across sections, it would be reasonable to hypothesize that linguistic features also vary across moves within a section, not just across sections. In fact, a few studies attempt to link choices of linguistic features with rhetorical structure at the move level. However, most of these studies are on the abstract genre rather than on the main research article. For instance, apart from the generic structure of the abstract, Lorés (2004) also investigates the thematic structure of the rhetorical moves; and Salager-Meyer (1992) examines the distribution of modality and verb tenses across the different moves of the abstract. Although the focus on certain linguistic features of interest in earlier works allows in-depth investigation, an in-depth description of the linguistic realizations of various moves in both the abstract and the main article would give the novice writer a more comprehensive view of the linguistic realizations of the article.

Previous studies have looked at move structures in various disciplines, but the attention focuses more on the "hard sciences" (i.e., natural sciences) than on the "soft sciences" (i.e., social sciences and humanities). Some of these studies also attempt to examine the variation of move structure across related genres. For example, Nwogu (1990) compares three genres: the abstract, the research article proper, and the popularized version of the research article; and Samraj (2005) compares the research article abstract with the introduction of the main research article. However, both studies are on natural sciences writing, and the focus is on rhetorical structure rather than the linguistic realizations of moves. It would be useful for novice writers to see what differences exist between the two seemingly similar yet distinct genres of research article abstracts and introductions.

The present study thus aims to explore the distribution patterns of various linguistic features in different moves in the introduction of research articles and their corresponding abstracts in applied linguistics and educational technology to identify the prototypical linguistic realizations of each move. Variations of these features across disciplines and genres are also discussed in this paper.

2. Methods

2.1 The construction of the corpus

A corpus of 40 empirical research articles was compiled from four journals in the fields of applied linguistics and educational technology: ten from the *Modern Language Journal* (MLJ) and ten from *TESOL Quarterly* (TQ) in the field of applied linguistics; ten from *Computers and Education* (CE) and ten from the *Journal of Computer Assisted Learning* (JCAL) in the field of educational technology. These journals were chosen because they have high impact factors according to Journal Citation Reports (2006).

First, I collected all the empirical research articles published in those four journals between January 2006 and May 2007. The inclusion of recent issues of journals ensures that the articles reflect the current trends. From this pool, ten articles were selected randomly from each journal.

2.2 The coding of rhetorical moves

The 40 articles were downloaded, converted from .pdf format into .txt format, and xml-tagged for move structure. The present study is part of a larger project examining the rhetorical structure and linguistic realizations of moves in the whole article, from the abstract through to the conclusion section, but this paper reports relevant findings only for the abstract and introduction of the articles. Therefore, the introductions of the articles and the corresponding abstracts were extracted for this particular study. Details of the four subcorpora are as follows:

Table 3. Frequency details of the four subcorpora.

Discipline	Abstracts			Introductions		
	texts	words	mean length	texts	words	mean length
Applied Linguistics	20	3,574	178.70	20	43,267	2,163.35
Educational Technology	20	3,449	172.45	20	32,631	1,631.55
Total	40	7,023	175.58	40	75,898	1,897.45

This study applied Swales's (1990, 2004) move structure concept. However, as the present study aimed at investigating the linguistic realizations of moves, the identification of moves was based only on textual function by using a detailed set of questions that coders can ask themselves (see Figure 1 below). The framework for the coding of moves in the abstracts and introductions (as shown in Figure 1) was based on Santos (1996) and Swales (2004). Although neither Santos's nor Swales's corpus includes Educational Technology articles, their models of move structure for abstracts and introductions seem to apply well to the data of the present study. No new functions had to be added to the framework.

	Function/description	Question asked
Abstract moves		
situating the research (STR)	setting the scene for the current research	What is known in the field?
presenting the research (PTR)	stating the purpose of the study, research questions and hypotheses	What is the study about?
describing the methodology (DTM)	describing the materials, subjects, variables, procedures, etc.	How was the research done?
summarizing the findings (STF)	reporting the main findings of the study	What did the researcher find?
discussing the research (DTR)	interpreting the results/findings and/or giving recommendations	What do the results mean?
Introduction moves		
establishing a territory (EST)	reviewing previous studies, leading into the present study	What has been done about the topic of research? What is the writer's view?
establishing a niche (ESN)	indicating a gap in previous research	What has not been done in the field?
presenting the present work (PPW)	announcing the purpose or content of the study	What is the purpose of the study? What are the research questions or hypotheses?

Figure 1. Framework for move coding in the present study.

After I completed the coding of the whole corpus, a research assistant independently coded the entire corpus again. A comparison of the two coding results yielded a high interrater reliability rate: *kappa* = 0.91 (using SPSS).

2.3 Approach to the analysis of linguistic realizations of moves

The linguistic features of moves listed below are based on those identified as characteristic of moves in a previous pilot study (Pho, 2008):

- Feature 1: Self-reference words
  Type 1: First person pronouns (e.g., *I, we, me, us*)
  Type 2: Other self-reference words (e.g., *the author*[*s*]*, the researcher*[*s*])
- Feature 2: Modal auxiliaries and semimodal verbs
  Type 3: Obligation/necessity modal verbs

Type 4: Permission/possibility modal verbs
Type 5: Volition/prediction modal verbs

- Feature 3: Verb tense and aspect
  Type 6: Present simple
  Type 7: Past simple
  Type 8: Present perfect
  Type 9: Past perfect
- Feature 4: Voice
  Type 10: Passive
  Type 11: Active
- Feature 5: Stance adjectives, adverbs, and nouns
  Type 12: Attitudinal stance words
  Type 13: Epistemic stance words
- Feature 6: *That*-complement clauses
  Type 14: *That*-complement clauses controlled by adjectives
  Type 15: *That*-complement clauses controlled by verbs
  Type 16: *That*-complement clauses controlled by nouns

After the corpus was xml coded for moves and steps, it was tagged for parts of speech using CLAWS7 (Garside and Smith, 1997). Using WordSmith (Scott, 2004) and Perl scripts, the distribution patterns of each linguistic feature in each rhetorical move were obtained. The concordances were manually checked to exclude cases that do not belong to the category (more details will be given in Section 3 below).

3. Findings and discussion

The number of occurrences of each linguistic feature in the five moves of the abstracts and the three moves in the introduction was counted separately and then normalized to 10,000 words to ensure that the numbers are comparable to each other as the total number of words varies from one move to another (The tokens reported in this study refer to this normalized frequency).

A negative binomial model[1] (using glm.nb in the MASS package (Venables and Ripley, 2008) in *R* [*R* Development Core Team, 2008]) was fitted to the data to estimate the effects of Type for the two Disciplines (see Figure 2 below). To determine whether there is a difference between the distribution pattern of types between the two disciplines, the Type:Discipline interaction was dropped from the model. An ANOVA test of the main effects fit and the effects after the interaction between Type and Discipline was dropped shows that the overall interaction between Discipline and Type in general is not significant ($p = 0.91$). However, Anscombe tests of the Q-Q plots of the difference between individual deviance residuals[2] show that there is a significant difference between the two disciplines for some combinations of Move and Type, for example Type

12 of Move 1 (STR), Types 4 and 5 of Move 2 (PTR), Types 9 and 15 of Move 3 (DTM), or Type 12 of Move 6 (EST).

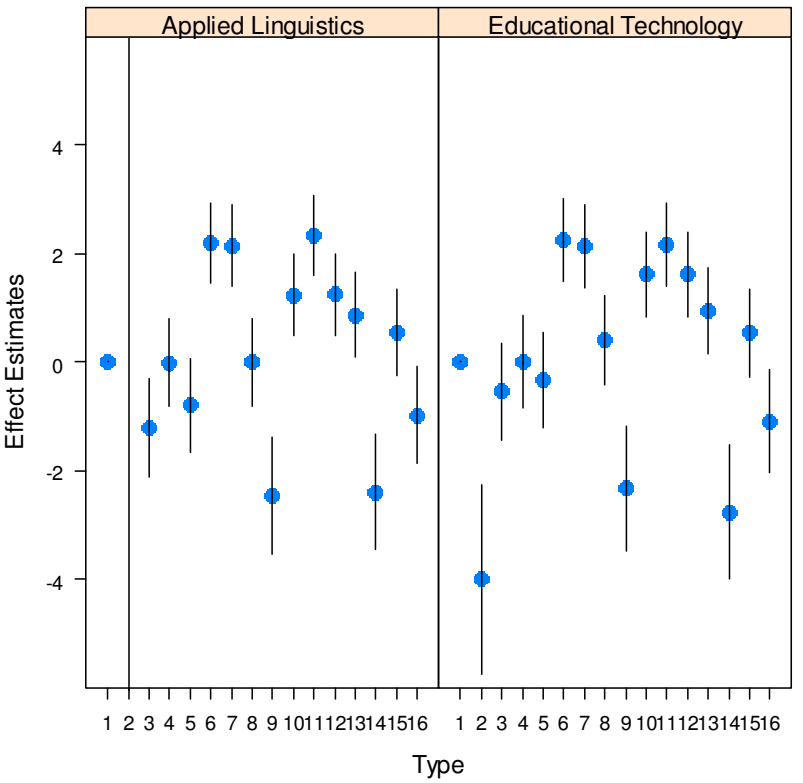


Figure 2. Effect estimates for Type from fitted model.

To identify the typical types for each move, the deviance residuals were plotted against Move and Type for the two disciplines (see Figure 3). The size of each symbol in the graph is proportional to the absolute value of the deviance residual. A circle indicates that the deviance residual is positive; a triangle indicates a negative deviance residual. For example, it can be seen that the deviance residual for Type 8 (Present Perfect), STR move in Applied Linguistics is large and positive. The observed count is 200, while the predicted count from the model is 42. In contrast, the deviance residual for Type 8, STF move in the same discipline is large and negative. The observed count is 0, while the predicted count is 35.45. The deviance residuals for the corresponding type and move in Educational Technology have similar patterns—positive for Type 8, STR move, but negative for the same type, STF move. The observed count and

predicted count in STR are 198 and 42.46, and those in STF are 0 and 43.94, respectively.

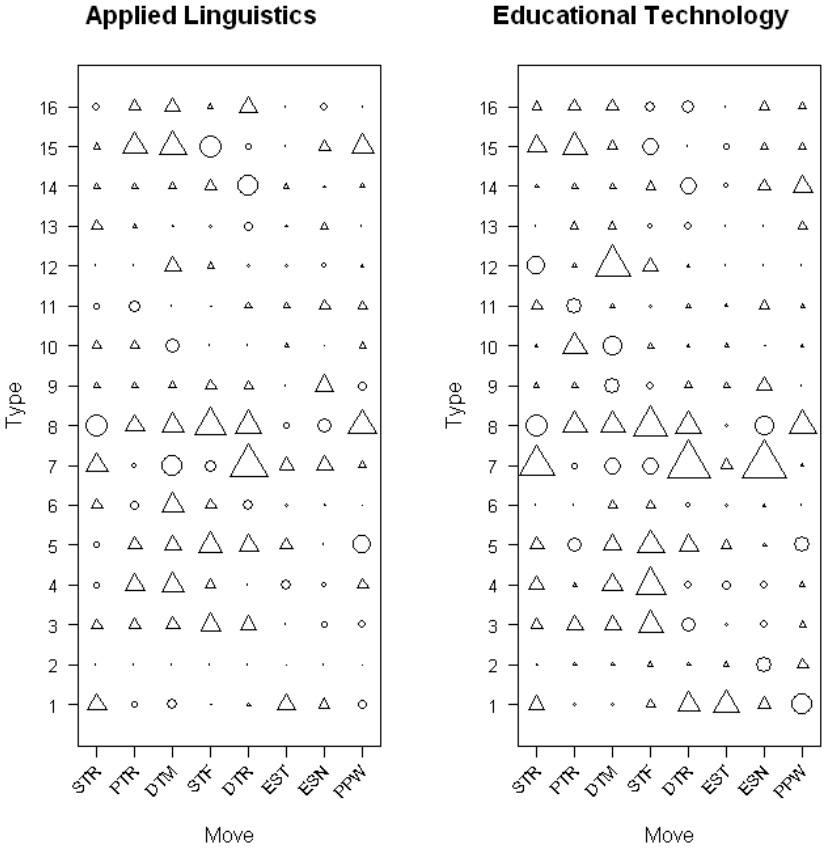


Figure 3. Deviance residuals from fitted model (Positive deviance residuals are indicated by circles, while negative deviance residuals are indicated by triangles, with the size proportional to the absolute value of the deviance residual.)

The subsections below will report the prototypical features of each move based on this figure (i.e., features that are distributed differently from other moves). Variation across disciplines will also be commented on wherever applicable.

3.1 Linguistic realizations of the abstract moves

3.1.1 Prototypical features of the *Situating the research* move

One of the linguistic features that makes the *Situating the research* move different from the other moves in the abstract is the distribution pattern of present perfect verbs, as mentioned in the example above. The distribution of this verb form in comparison with the other tenses can be seen in Figure 4:

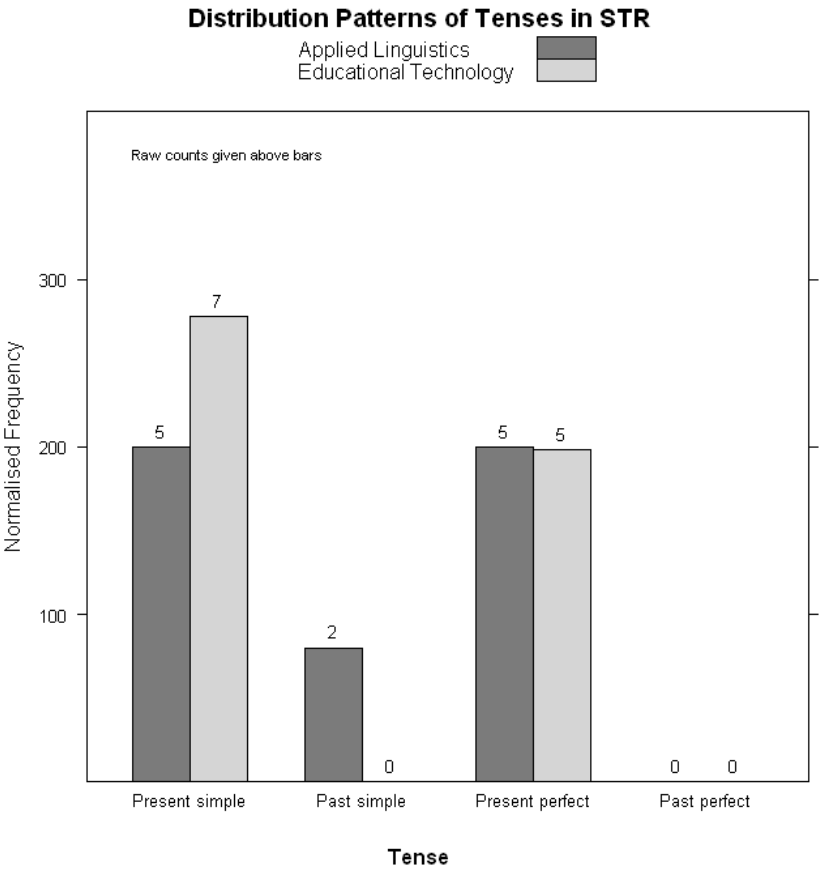


Figure 4. Distribution patterns of tenses in the *Situating the research (STR)* move.

The STR (*Situating the research*) move normally occurs at the beginning of the abstract, where the author gives a general view of previous research; therefore, it is not surprising that the present perfect is dominant in this move, together with the present simple. The following examples show how these two verb forms are used in the corpus:

(1) ***<STR>*** Although many studies have described the L2 learning opportunities created by individual tasks, considerably less research has investigated task-based syllabi and courses (Bruton, 2002; Candlin, 2001; Ellis, 2003; Skehan, 2003). [AL_TQ10]

(2) ***<STR>*** Gender differences in the pursuit of technology careers are a current issue of concern. [ET_JCAL10]

Unlike the other moves of the abstract, no self-reference words were found in the *Situating the research* move in either of the disciplines. Authors do reveal their own voice or judgment in this move through other means, which vary across the disciplines. As can be seen in Figure 3, while applied linguistics authors tend to use modal auxiliaries and semimodal verbs (particularly permission/possibility and volition/prediction modal verbs) in situating their study, educational technology authors use a great number of attitudinal stance words (the observed count is 437 while the predicted count is only 141.67). Typical cases are given below:

(3) ***<STR>*** In designing suitable listening tests, teachers can provide various forms of support to reduce the demands of the task for the test takers. [AL_TQ6]

(4) ***<STR>*** Adapting training methods to specific teacher traits to best facilitate the training effects for preservice teachers is an important, yet neglected, topic in aptitude-treatment interaction research. [ET_CE9]

3.1.2 Prototypical features of the *Presenting the research* move

Like the *Situating the research* move, the combination of verb tense and aspect is also a feature that distinguishes the *Presenting the research* (PTR) move from the other moves in the abstract. This is the only move in the abstract where the deviance residuals for both past simple and present simple verbs are both positive (see Figure 3 above), and this applies to both disciplines. The observed counts for present simple (Type 6) and past simple (Type 7) verbs in Applied Linguistics are 260 and 214, respectively, while the predicted counts are 15.42 and 148.09. Similarly, the observed counts for these two verb types in Educational Technology are 208 and 236 for observed counts and 17.5 and 15.69 for predicted counts, respectively. Almost all the authors begin this move with either *this article, this paper,* or *this study* as a way of introducing their own study. Interestingly, when the subject of the sentence is *this article* or *this paper,* the verb is always in the present simple. When the subject is *this study,* the verb is in past simple most of the time. Two typical examples are given below:

(5) ***<PTR>*** *This paper* analyses a sample of online discussions to evaluate the development of adult learners as reflective practitioners within a networked learning community. [ET_JCAL9]

(6) ***<PTR>*** *This study* investigated the effects of four types of listening support: previewing the test questions, repetition of the input, providing

background knowledge about the topic, and vocabulary instruction. [AL_TQ6]

Another feature which is also typical of this move is the predominant use of active verbs.[3] The observed counts for active verbs (Type 11) are 366 for Applied Linguistics and 389 for Educational Technology, while the predicted counts for this type in the two disciplines are only 17.86 and 15.97, respectively. As can be seen in Examples (5) and (6) above, the authors often present their study by using an action verb in active voice, which makes the *Presenting the research* move different from the other moves as discussed later in the other sections below.

3.1.3 Prototypical features of the *Describing the methodology* move

As can be seen from Figure 3, what makes the *Describing the methodology* (DTM) move stand out from the other moves is the positive deviance residuals of passive verbs (Type 10). The observed counts for this type are higher than the predicted counts in both disciplines—167 vs. 72.71 in Applied Linguistics, and 257 vs. 86.91 in Educational Technology. The preference of passive verbs over active verbs, as shown in the following extract, can be attributed to the fact that authors try to stay as objective as possible when they present the methods of their study.

(7) ***<DTM>*** Survey data were collected from 922 students in 51 courses at both the graduate and undergraduate levels. [ET_CE8]

Another feature that is also characteristic of the *Describing the methodology* move is the use of past simple verbs. The deviance residuals for Type 7 in both disciplines are large and positive. The observed counts are 548 for Applied Linguistics and 385 for Educational Technology, whereas the predicted counts for these two disciplines are only 18 and 14.76, respectively. The frequent use of past tense verbs in this move is understandable, as here the authors describe how the data were collected and analyzed before the write-up of the paper. Past simple almost always goes hand in hand with passive voice for most of the verbs of this move, as can be seen in Example (7) above. Past tense also occasionally occurs with active voice, as in Example (8) below, but most of these verbs are intransitive verbs, as in Example (9), and thus, as mentioned before, not counted as "active verbs" in the present study.

(8) ***<DTM>*** We examined the validity of 2 types of assessments: an off-task self-assessment and an on-task self-assessment. [AL_MLJ10]

(9) ***<DTM>*** The study was a longitudinal qualitative case study in one faculty at a large North American university. [AL_TQ5]

Unlike the other abstract moves, no instances of modal verbs were found in the *Describing the methodology* move of either discipline. The deviance residuals for all the three types of modal verbs (Types 3, 4, and 5) are negative.

3.1.4 Prototypical features of the *Summarizing the findings* move

Like the *Describing the methodology* move, the verbs in the *Summarizing the findings* (STF) move are typically in past simple (Type 7). The observed count for this type in Applied Linguistics is 568, while the predicted count is 30.45; the figures for Educational Technology are 606 and 25, respectively, One example from the corpus is given below:

(10) **<*STF*>** The results of the study indicated that the intensive use of ICT and the process-oriented learning environment supported the development of student expertise. [ET_CE2]

While authors seldom use modal auxiliaries and semimodal verbs in the *Summarizing the findings* move (as can be seen from the negative deviance residuals for Types 3, 4, and 5, STF move in Figure 3), they tend to use stance words, especially epistemic stance words, compared to the first three moves reported above. This trend is similar for both disciplines. The use of epistemic stance words, as exemplified in the following extract, shows that authors try to avoid overgeneralizations of their findings:

(11) **<*STF*>** It was also found that the on-task self-assessment was generally less influenced by student attitude/personality factors than was the off-task self-assessment. [AL_MLJ10]

Another linguistic feature that stands out in this move is the use of *that*-complement clauses. The deviance residual for Type 15 (*that*-complement clauses controlled by verbs) is large and positive for both disciplines, as illustrated in the following example:

(12) **<*STF*>** The findings *revealed* that learners made significant improvements in both content knowledge and functional linguistic abilities. [AL_MLJ9]

It seems that this structure helps authors project their findings more easily by signaling that the move is now changed to the reporting of the findings in their own study. The deviance residual for Type 16 (*that*-clauses controlled by nouns) is also positive in Educational Technology, but it is negative in Applied Linguistics (the expected count is 13.4, while the observed count is eight).

3.1.5 Prototypical features of the *Discussing the research* (DTR) move

As can be seen in Figure 3, the most frequent feature in this move is *that*-complement clauses. However, there is a difference in the distribution of *that*-

clause types across the two disciplines. Type 14 (*that*-clause controlled by adjectives) and, to a lesser extent, Type 15 (*that*-clause controlled by verbs) are typical of this move in Applied Linguistics, whereas Types 14 and 16 are typical features in Educational Technology. The following examples illustrate the use of *that*-clauses controlled by verb and noun:

(13) ***<DTR>*** We *suggest* that using the keyword method with phonological keywords and direct L1 keyword-translation links in the classroom leads to better L2 vocabulary learning at early stages of acquisition. [AL_MLJ3]

(14) ***<DTR>*** The key *conclusion* of the study is that integration may be a desirable option regardless of the potential extra costs involved. [ET_CE4]

The deviance residual for Type 13 (epistemic stance adjectives, adverbs, and nouns) is also positive for both disciplines. The observed count is 189, while the predicted count is 104.3 for Applied Linguistics. Similarly, 176 tokens were observed in Educational Technology while only 105.54 are expected. This is not surprising as the authors normally interpret their findings in this move and in doing so they add in words that help avoid making overgeneralizations of the results, as illustrated in the following extract:

(15) ***<DTR>*** It is possible that computerised assessment does not detect the established gender effect due to differences between males and females in motivation, computer experience, and competitiveness. [ET_JCAL7]

One of the features that make the *Discussing the research* move different from the *Summarizing the research* move in particular or other abstract moves in general is the use of modal auxiliaries and semimodal verbs, especially those referring to obligation/necessity or possibility. For example,

(16) ***<DTR>*** This study suggests that recasts vary in implicitness and that these differences may have an impact on their effectiveness, both in terms of learners' successful uptake and subsequent use. [AL_MLJ2]

Finally, one distinctive feature of the *Discussing the research* move is the dominant use of present simple verbs. Both of the deviance residuals for Type 6 (present simple) are positive in both disciplines. The observed counts for this type are 706 in Applied Linguistics and 554 in Educational Technology, whereas the corresponding predicted counts are only 38.9 and 39.29. Such a distribution pattern of verb tense and aspect is very different from those in the *Describing the methodology* move and the *Summarizing the findings* move that I reported above, with the majority of verbs in past simple.

3.2 Linguistic realizations of the introduction moves

3.2.1 Prototypical features of the *Establishing a territory* move

The distribution patterns of linguistic realizations of the *Establishing a territory* (EST) move in the introduction section of the main article are more similar to those of the *Establishing a niche* (ESN) move than to those of the *Presenting the present work* (PPW) move. Both have negative deviance residuals for Type 1 (first person pronouns such as *I* or *we*). This result can be expected as the author is focusing on other researchers' work in this move. Even when the authors interpret or discuss others' findings, they tend to avoid the direct subject *I* or *we*.

In contrast, the deviance residuals for Type 15 (*that*-complement clauses controlled by verbs) are positive in both disciplines. The observed counts are 80 for Applied Linguistics and 84 for Educational Technology, whereas the expected counts are 0.68 and 0.59, respectively. These structures are generally used to report other researchers' findings or arguments, as shown in the following extracts:

(17) **<*EST*>** Morahan-Martin (1999) *found* that women college students went online less frequently, spent less time per session, and used the Internet for fewer purposes than men. [ET_CE5]

(18) **<*EST*>** Such findings *suggest* that literacy promotes awareness of linguistic segments in oral language processing. [AL_TQ2]

Modal verbs can also be said to be characteristic of this move. Although all three categories of modal verbs were found, Type 4 (permission/possibility modal verbs) is the most frequently used. Authors tend to use such modal verbs as *can* or *may* as shown in the following examples in their discussion or interpretation of other studies:

(19) **<*EST*>** We can suggest that girls' ICT competence increases with time, and that they may reach a high level of understanding of and competence in, e.g., communication-related applications. [ET_CE2]

(20) **<*EST*>** The absence of L2 proficiency from the model, however, suggests that L2 proficiency may not be an influential factor. [AL_TQ7]

3.2.2 Prototypical features of the *Establishing a niche* move

As mentioned in the previous section, the *Establishing a niche* move shares some features with the *Establishing a territory* move, for example, the low frequency of self-reference words or the relatively high frequency of modal verbs. However, there are features that distinguish between these two moves. As shown in Figure 3, the deviance residuals for Type 8 (present perfect) are large and positive for both disciplines. The observed counts are 92 and 198, whereas the expected counts are only 4.05 and 6.94 in Applied Linguistics and Educational

Technology, respectively. This verb form seems to help the author indicate a gap in previous research more easily with a subject referring to past studies in general, as in the following extract:

(21) ***<ESN>*** Much of the research on interlanguage pragmatics (ILP) has focused on language use by second language (L2) learners; in other words, their production of target language speech acts (SAs), rather than on the development of their pragmatic competence (Kasper, 1996; Kasper and Rose, 1999, 2002). [AL_MLJ1]

Another type that has positive deviance residuals in this move is Type 12 (attitudinal stance words). The observed count in Applied Linguistics is 193, while the expected count is 14.08. Similarly, the observed count in Educational Technology is 271, whereas the predicted count is 23.25. A closer look at the attitudinal stance words in this move revealed that, unlike the *Establishing a territory* move, most of the attitudinal stance words in the *Establishing a niche* move are negative words. The following extracts show some of the most typical negative words found in this move:

(22) ***<ESN>*** Specifically, there is limited empirical evidence to date supporting a positive impact on student learning and students' and professors' perceptions of the classroom experience. [ET_CE3]

(23) ***<ESN>*** … relatively few empirical studies have documented how teachers and learners react to entirely task-based courses, as opposed to the use of individual task … [AL_TQ10]

3.2.3 Prototypical features of the *Presenting the present work* move

Like the *Establishing a niche* move, there is not much use of the *that*-complement structures in *Presenting the present work* move. The deviance residuals for Types 14, 15, and 16 are negative. Although the total number of modal verbs in this move is about the same as in the other two moves, what makes this move stand out is the use of volition-prediction modal verbs (as indicated by the positive deviance residuals for Type 5 in Figure 3 above)—51 tokens observed in Applied Linguistics vs. 1.9 expected and 60 observed vs. 2.48 expected in Educational Technology. Most of the volition-prediction modal verbs are used to introduce hypotheses of the study:

(24) ***<PPW>*** It was expected that children would consider fewer errors acceptable when the target was presented close to distracting objects and that they would need more time to attain this high level of accuracy. [ET_CE6]

(25) ***<PPW>*** Specifically, the research addressed the following questions: Will different types of listening support affect learners' listening performance differently? [AL_TQ6]

The most noticeable characteristic of the PPW move is the use of self-reference words. The deviance residuals for Type 1 are positive for both disciplines. The most popular self-reference word used in this move is *we*. This can be explained by the fact that the majority of articles in the corpus were written by more than one author:

(35) ***<PPW>*** We expected that students at different levels of computer literacy differ with respect to the patterns of media use in the computer-based learning environment processes. [ET_JCAl4]

4. Conclusion

The findings of this study show that linguistic features vary more across moves than disciplines. The same move in two different disciplines can have similar distribution patterns of a certain linguistic feature. For example, Type 8 (present perfect) is typical for the STR move in both disciplines, or Type 5 (volition/prediction modal verbs) is positive for the *Presenting the present work* move across the two disciplines. The similarity between the two disciplines can be attributed to their being multidisciplinary and their belonging to the same broad field of teaching and learning, although one has a focus on the linguistics aspect and the other on technology.

The variation of linguistic features across moves demonstrated in the present study indicates that it would be an overgeneralization to simply state that certain features are typical of the whole section of an article as some previous studies have done. Linguistic features do vary across moves, not just sections as a whole. This is understandable, since different linguistic forms serve different textual functions. However, it should be noted that a move is realized by a cluster of linguistic features rather than a single feature. Thus, for example, we cannot say that a high frequency of *that*-complement structures in an abstract move will decide that it is the *Discussing the research* move. We have to take into consideration other features such as verb tense, modal verbs, and stance words.

The characterizations of each move should give novice writers a more comprehensive and specific view of how they can write an abstract and introduction of an article to be published in an English-language journal.

5. Notes

* I am grateful to the reviewers for their useful comments. I would also like to thank Dr. Julie Bradshaw, Prof. Kate Burridge, and Dr. Simon Musgrave for reading earlier drafts of this paper.

1. The model included terms for Type, Move, and Discipline, and allowed interaction between Type and Discipline and between Move and Discipline.

2. The Deviance Residuals (see, for example, Venables and Ripley [2002, p. 189]) measure the discrepancy of the fitted model from the data and is used to determine the interaction effect between Type and Move.
3. Note that the passive verbs in the present study are compared only to active verbs that are transitive, since only those verbs have the true potential of being used in either active or passive voice.

References

Burrough-Boenisch, J. (2002), "Examining present tense conventions in scientific writing in the light of reader reactions to three Dutch-authored discussions," *English for specific purposes,* 22(1): 5–24.

Burrough-Boenisch, J. (2003), "Shapers of published NNS research articles," *Journal of second language writing,* 12(3): 223–243.

Cho, S. (2004), "Challenges of entering discourse communities through publishing in English: perspectives of nonnative-speaking doctoral students in the United States of America," *Journal of language, identity, and education,* 3(1): 47–72.

Flowerdew, J. (1999), "Problems in writing for scholarly publication in English: the case of Hong Kong," *Journal of second language writing,* 8(3): 243–264.

Garside, R., and N. Smith (1997), "A hybrid grammatical tagger: CLAWS4," in: R. Garside, G. Leech, and A. McEnery (eds.) *Corpus annotation: linguistic information from computer text corpora*. London: Longman, 102–121.

Gosden, H. (1992a), "Discourse functions of marked theme in scientific research articles," *English for specific purposes,* 11(3): 207–224.

Gosden, H. (1992b), "Research writing and NNSs: from the editors," *Journal of second language writing,* 1(2): 123–139.

Graetz, N. (1985), "Teaching EFL students to extract structural information from abstracts," in: J. M. Ulijn and A. K. Pugh (eds.) *Reading for professional purposes: studies and practices in native and foreign languages*. London: Heinemann Educational Books, 123–135.

Harwood, N. (2005), "'Nowhere has anyone attempted ... In this article I aim to do just that'": a corpus-based study of self-promotional *I* and *we* in academic writing across four disciplines," *Journal of pragmatics,* 37(8): 1207–1231.

Hood, S. (2005), "What is evaluated, and how, in academic research writing? The co-patterning of attitude and field," *Australian review of applied linguistics*, 19: 23–40.

Hyland, K. (1996), "Talking to the academy: forms of hedging in science research articles," *Written communication,* 13(2): 251–281.

Hyland, K. (2002), "Activity and evaluation: reporting practices in academic writing," in: J. Flowerdew (ed.) *Academic discourse*. Harlow: Pearson Education Limited, 115–130.

Kuo, C.-H. (1999), "The use of personal pronouns: role relationships in scientific journal articles," *English for specific purposes,* 18(2): 121–138.

Lillis, T., and M. J. Curry (2006), "Professional academic writing by multilingual scholars: interactions with literacy brokers in the production of English-medium texts," *Written communication,* 23(1): 3–35.

Lorés, R. (2004), "On RA abstracts: from rhetorical structure to thematic organisation," *English for specific purposes,* 23(3): 280–302.

Malcolm, L. (1987), "What rules govern tense usage in scientific articles?" *English for specific purposes,* 6(1): 31–43.

Martínez, I. A. (2001), "Impersonality in the research article as revealed by analysis of the transitivity structure," *English for specific purposes,* 20(3): 227–247.

Martínez, I. A. (2003), "Aspects of theme in the method and discussion sections of biology journal articles in English," *Journal of English for academic purposes,* 2(2): 17–37.

Misak, A., M. Marusic, and A. Marusic (2005), "Manuscript editing as a way of teaching academic writing: experience from a small scientific journal," *Journal of second language writing,* 14(2): 122–131.

Nwogu, K. N. (1990), *Discourse variation in medical texts: schema, theme and cohesion on professional and journalistic accounts*. Nottingham: Department of English Studies, University of Nottingham.

Pagel, W. J., F. E. Kendall, and H. R. Gibbs (2002), "Self-identified publishing needs of nonnative English-speaking faculty and fellows at an academic medical institution," *Science editor,* 25(4): 111–114.

Pho, P. D. (2008), "Research article abstracts in applied linguistics and educational technology: a study of linguistic realizations of rhetorical structure and authorial stance," *Discourse studies,* 10(2): 231–250.

Posteguillo, S. (1999), "The schematic structure of computer science research articles," *English for specific purposes,* 18(2): 139–160.

R Development Core Team (2008), R*: a language and environment for statistical computing*. Vienna, Austria: *R* Foundation for Statistical Computing.

Salager-Meyer, F. (1992), "A text-type and move analysis study of verb tense and modality distribution in medical English abstracts," *English for specific purposes,* 11(2): 93–113.

Samraj, B. (2002), "Introductions in research articles: variations across disciplines," *English for specific purposes,* 21(1): 1–17.

Samraj, B. (2005), "An exploration of a genre set: research article abstracts and introductions in two disciplines," *English for specific purposes,* 24(2): 141–156.

Santos, M. B. D. (1996), "The textual organization of research paper abstracts in applied linguistics," *Text*, 16(4): 481–499.

Scott, M. (2004), *Wordsmith Tools 4.0*. Oxford: Oxford University Press.

Sionis, C. (1995), "Communication strategies in the writing of scientific research articles by non-native users of English," *English for specific purposes,* 14(2): 99–113.

Swales, J. (1990), *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Swales, J. (2004), *Research genres: explorations and applications*. New York: Cambridge University Press.

Tarone, E., S. Dwyer, S. Gillette, and V. Icke (1998), "On the use of the passive and active voice in astrophysics journal papers: with extensions to other languages and other fields," *English for specific purposes,* 17(1): 113–132.

Thompson, G., and Y. Ye (1991), "Evaluation in the reporting verbs used in academic papers," *Applied linguistics,* 12(4): 365–382.

Tucker, P. (2003), "Evaluation in the art-historical research article," *Journal of English for academic purposes,* 2(4): 291–312.

Vassileva, I. (2001), "Commitment and detachment in English and Bulgarian academic writing," *English for specific purposes,* 20(1): 83–102.

Venables, W. N., and B. D. Ripley (2002), *Modern applied statistics with* S. New York: Springer.

Venables, W. N., and B. D. Ripley (2008), *Mass: main package of Venables and Ripley's modern applied statistics with* S.

Ventola, E. (1994), "Abstracts as an object of linguistic study," in: S. Cmejrkova, F. Danes, and E. Havlova (eds.) *Writing vs. speaking: language, text, discourse, communication*. Tubingen: Günther Narr, 333–352.

Yang, R., and D. Allison. (2004), "Research articles in applied linguistics: structures from a functional perspective," *English for specific purposes,* 23(3): 264–279.

Lexical bundle distribution in university classroom talk

Eniko Csomay

San Diego State University

Viviana Cortes

Georgia State University

Abstract

The present study investigates the relationship between the discourse functions of lexical bundles found in classroom teaching and their position. Eighty-four lexical bundles, frequently occurring four-word combinations identified earlier in university classroom talk (Biber, Conrad, and Cortes, 2004), are tracked in the first six Vocabulary-Based Discourse Units (VBDUs) also identified previously (Biber, Csomay, Jones, and Keck, 2004) of 176 university lectures. Among others, expressions such as you might want to, I would like to, if you look at, *and* in the case of *are traced in tandem with their previously identified classification of discourse functions. While earlier studies reported on the relationship between the bundles' discourse functions and their position in the first three discourse units (Cortes and Csomay, 2007), there are no studies yet on how the frequency patterns may change in the second set of three discourse units.*

The findings of this study show a sharp increase in the use of referential bundles and those discourse organizers with a topic elaboration that focuses in the second set of discourse units. At the same time, the use of bundles expressing stance, especially those referring to personal ability and personal intention and those discourse organizers with a topic introduction, drop in the second set of discourse units. These findings provide further, lexical evidence for the claim that a strong relationship exists between intra-textual linguistic variation and the corresponding shift in discourse functions in university classes (Csomay, 2005, 2007).

1. Introduction

Among the growing number of research applying corpus-based methodologies to describe academic language use are comprehensive linguistic descriptions of spoken and written registers in the academic context (Biber, Conrad, Reppen, Byrd, and Helt, 2002; Biber, 2006), analyses of individual lexico-grammatical items (e.g., pronouns by Fortanet, 2004), or analysis of frequently occurring four-word combinations (lexical bundles) as they appear in academic resisters (Biber, Conrad, and Cortes, 2004; Nesi and Bastrukmen, 2006). Another area of research provides insights about the relationship between linguistic variation and discourse structure (Biber, Connor, Csomay, Jones, Keck, and Upton, 2007). Although corpus-based studies are growing in number in all of these areas, so far only one study applied corpus-based methods to seek the relationship between the position

of lexical bundles mentioned above and discourse structure (Cortes and Csomay, 2007). The present study elaborates on this study in that it looks at variation in the use of lexical bundles in two sets of discourse units: the initial three units are compared to the second three units in terms of their lexical bundle make-up.

Lexical studies of university classroom talk were conducted by researchers from two different approaches to text analysis: top-down approaches usually take predefined units of analysis, in this case lexical categories, in search of variation of functional correlates, while bottom-up approaches look at the lexical patterns as they emerge from the text first and then interpret the findings (Biber, Connor, and Upton, 2007).

Corpus linguistic research applying top-down analyses on lectures has grown since 2001, when the Michigan Corpus of Academic Spoken English (MICASE) was introduced and made available online. Scholars have discussed the use of reflexivity (Mauranen, 2001), the use and functions of evaluative adjectives (Swales and Burke, 2003) and pronouns (Fortanet, 2004), and the kinds of idioms used in the classroom (Simpson and Mendis, 2003). Investigations on MICASE identify discourse markers that mark new episodes (Swales and Malczewski, 2001) or compare discourse markers in MICASE with those in another corpus of guest lectures (Camiciottoli, 2004). Similar to earlier analyses, these studies used previously identified lexical items.

Another group of scholars took a bottom-up approach to describe the linguistic patterns of classroom talk as they emerge in a subcorpus of the TOEFL 2000 Spoken and Written Academic Language corpus (T2KSWAL) (Biber et al., 2002; Biber, 2006). They, for example, characterized the language of university classrooms in terms of the co-occurring lexico-grammatical patterns and compared the findings to those in other academic registers, such as textbooks (Biber et al., 2004), and non-academic registers, such as conversation (Biber et al., 2002; Csomay, 2006). As for lexical studies in this school of thought, computational tools were developed to identify two different types of units of analysis in discourse: (a) automatic text segmentation techniques were developed (Biber, Csomay, Jones, and Keck, 2004) and applied to extract empirically based discourse units in classroom talk, and (b) automatic identification of the most frequently occurring four-word sequences (Biber, Conrad, and Cortes, 2004) was developed to extract these lexical patterns in classroom talk. After identifying these new units of analysis, both groups applied further analytical techniques to provide linguistic characterizations. More specifically, the lexico-grammatical features of the discourse units mentioned above were described in order to depict the relationship between linguistic variation within text and macro-structural discourse functions (Csomay, 2005, 2007). The structural descriptions of the four-word sequences, called lexical bundles, and their functional categories in classroom talk have also been described.

This study looks for associations between patterns of discourse structure and lexical bundle functions. First, we introduce the discourse segmentation technique, followed by the way lexical bundles have been identified and categorized.

1.1 Vocabulary-Based Discourse Units and discourse structure

Vocabulary-Based Discourse Units (VBDUs) are lexical episodes in discourse.[1] These lexical episodes are present in any text and are identified as the interplay between repeated vocabulary and vocabulary newly entering a stretch of discourse (Youmans, 1991). Following this basic principle, a computer program was developed to automatically track these patterns in discourse (Biber, Connor, Csomay, Jones, Keck, and Upton, 2007) and, based on a set of algorithms, to automatically segment the discourse into well-defined subunits (Csomay, 2002) depicting this change. In this study, we used a modified version of a computer program that includes a computational procedure called TextTiling, coined by Hearst (1997). As described in multiple earlier studies (Biber, Csomay, Jones, and Keck, 2004; Biber, Connor, Csomay, Jones, Keck, and Upton, 2007; Cortes and Csomay, 2007), the program tracks these alternating patterns in adjacent segments of text. As Csomay (2005: 247) indicates,

> The text is processed via a "sliding window" of 100 words, and the program compares the first 50 words in that window to the second 50 words. That is, at the start, the window is positioned at the beginning of the text and contains words 1–50 in the first half of the window and 51–100 in the second half. Then the window "slides" one position and contains words 2–51 in the first half and 52–101 in the second half. The window continues to slide one position at a time, allowing the comparison of two 50-word chunks of the target text, until the end of the text is reached. In the meantime, at each word, a similarity value is calculated (see Hearst 1991, 1994 for details) that indicates the extent to which the words in the first half of the window are identical within those in the second half. If the two halves use the same vocabulary to a large extent, they are considered to belong to a single VBDU and are interpreted as lexically coherent units (Biber, Csomay, Jones, and Keck, 2004). In contrast, when the two segments are maximally different in their vocabulary, they are considered to mark the boundaries between two VBDUs.

The accumulation of new vocabulary into the discourse is claimed to introduce new topics (Prince, 1981; Youmans, 1991). The similarity value calculated at each word indicates the extent to which the vocabulary in the two adjacent segments is identical (orthographically). For example, a value of 0.25 means a 25% overlap. If the similarity value is low, it shows less similarity in the lexical patterns of the two segments; hence, in Prince's terms, a change in topic.

Using this methodology, subsequent studies segmented discourse into these subunits, VBDUs, and applied multidimensional analyses to characterize their linguistic make-up, which ultimately led to descriptions of linguistic variation within texts (Csomay, 2005, Biber, Connor, Jones, and Upton, 2007). These studies have also shown how linguistic variation in these units relates to variation in discourse functions in the macro-structure of discourse.

Specific to spoken discourse in the classroom, earlier studies applying the analytical framework discussed above have shown that the first three units in university classroom talk exhibit different linguistic features from the subsequent units (Csomay, 2005). Corresponding to the change in the linguistic characteristics of these two sets, there is a shift in the communicative and instructional purposes between the first three units and the subsequent unit(s). More specifically, the first three VBDUs exhibited a constellation of turn-taking patterns and linguistic features that reflects discourse associated with classroom management type of functions. In contrast, the subsequent units showed co-occurring patterns of turns by single speakers (monologic talk) and linguistic features associated with an informational focus and conceptual orientation. As described in linguistic terms, the first set of units (of three) was very different from subsequent units. To grant comparability, in this study, we take the first three units and compare them to the second three units for lexical bundle use.

1.2 Lexical bundles and university classroom talk

Lexical bundles are frequently occurring, four-word lexical sequences in a register (Biber et al., 1999) with no inherent structural integrity. Although at times they might resemble well-known fixed expressions, they are not. In everyday conversation, these expressions are, for example, *what do you mean, I don't know why,* and in academic prose*, as a result of, in the case of,* and *on the other hand*. Among others, in university lectures, we find expressions such as *if you look at, nothing to do with,* and *I want you to.*

Lexical bundles are identified automatically as a computer program stores every n-word sequence (also called n-grams) in a corpus. In terms of identifying n-grams, the computer program slides the text in a corpus word by word through an n-slot window, where one word fits in each slot of the window. At each word, the program takes a snapshot of the given n-word sequence and puts each snapshot in a database. As the text is sliding through this window and a snapshot is taken of each n-word sequence, the program also checks whether the latest sequence is already in the database. If not, it enters it; if yes, it adds one to the frequency count of that sequence. Although the identification happens across sentence and text boundaries, the program keeps a record of the number of texts each sequence occurs in. Clearly, n-grams can be combinations of two, three, or more words. For this study, only four-word sequences were dealt with since they have been found to include many three-word bundles in them (*as a result* is included in *as a result of*) and because four-word bundles, in general, are much more frequent than five-word bundles (Biber and Conrad, 1999; Cortes, 2004). While "the actual frequency cut-off used to identify lexical bundles is somewhat arbitrary" (Biber, Conrad, and Cortes, 2004), most studies identify lexical bundles as four-word sequences that occur at least a minimum of ten times in a million words in a corpus (Biber et al., 1999).

As mentioned before, these frequency cut-off points are not determined in relation to text boundaries, however: each bundle has to occur "in at least five different texts, to avoid idiosyncrasies" (Cortes and Csomay, 2007: 60).

Following this definition, Biber et al. (2004: 418) reported on four frequency bands for lexical bundles in university classroom teaching: 10–19 per million; 20–39 per million; 40–99 per million; and over 100 per million. In the present study, we report on the distribution patterns of only those bundles that appeared at least 40 or more times in a million words and as originally identified by Biber et al. (2004). We take the two bands with highest frequency simply because we look for bundle distribution in a specific and relatively small portion of the entire corpus in which the bundles were originally identified. We assume that bundles with higher overall frequency may have a wider distributional pattern and, hence, may also have better chances to occur in smaller portions of the corpus as well.

The methodology to extract bundles and the criteria to decide on the length and cut-off point of the bundles described above guided earlier studies that identified and listed those lexical bundle sequences in the class sessions that we used for the present study. Finally, lexical bundles are classified according to their structural make-up and according to the discourse functions they perform. The lexical bundles in classroom teaching fell into three major functional categories: discourse organizers, stance expressions, and referential expressions.

1.3 The goal of this study

The goal of the present study is to combine discourse pattern studies with lexical pattern studies. More specifically, the present study investigates the relationship between the discourse functions of lexical bundles found in classroom teaching (Biber, Conrad, and Cortes, 2004) and their position in a stretch of discourse. To achieve this goal, lexical bundles are tracked in the discourse units identified previously (Biber, Csomay, Jones, and Keck, 2007) in a set of university lectures.

The methodology to segment discourse into smaller units was described above (1.1), followed by the introduction of lexical bundles traced in those units (1.2). Next, we describe the computer programs that we developed to track lexical bundles in these units (2), followed by the presentation of findings (3), and the conclusions (4).

2. Methodology

In this study, we examine the first six discourse units (VBDUs) described in section 1.1 to look for the distribution patterns of the previously identified lexical bundles introduced in 1.2.

The VBDUs were identified in the class sessions of the T2KSWAL corpus (see Biber et al., 2002), constituting a total of 1,056 units. As for lexical bundles, we took a conservative approach and considered only those lexical bundles (84 in total) that occurred with an overall frequency cut-off point of 40 (or more) times in a million words and in five or more texts previously identified in the full class sessions of the T2KSWAL corpus (Biber, Conrad, and Cortes, 2004; Biber, 2006). Appendix A presents a list of these bundles, grouped by function. Overall, lexical bundles can be classified into three categories used for their functions in

classroom discourse: discourse organizers, referential expressions, and stance expressions.

In order to track the three aspects of bundles (type of function, bundle token, and bundle position), several computer programs were developed to perform the following steps:

- Retrieve all hits of the 84 bundles identified (earlier) in the six units.
- Note the bundles' positions in terms of VBDU number.
- Count each bundle in each position.
- Keep track of the functional category of each as classified earlier.
- Compute change: the percent difference between the frequencies of the first three versus the second three VBDUs.

Conceptually, change is perceived as the difference in frequency between the first and second set of VBDUs viewed from the perspective of the first set. Hence, it is calculated by deducting the frequency counts from the second set of VBDUs (i.e., units 4–6 [$freq_j$]) from the frequency counts from the first set of VBDUs (i.e., units 1–3 [$freq_i$]), divided by the frequency values in the first unit and multiplied by 100. To calculate change in the functional categories, then, first, the total frequencies of the bundles in each category were added and change was computed using the following formula.

$$\text{change} = \frac{freq_i - freq_j}{freq_i} \; x\ 100$$

For example, the total frequency of all bundles under the category *Stance markers* appear 225 times in set 1 (VBDU 1–3) and 181 times in set 2 (VBDU 4–6). Hence, using the formula above, we would get a 19.55% decline as we compare the first and second set (see Table 1 below). The direction of change is noted with an arrow pointing up or down, depending on whether the frequencies in the second set are higher or lower than in the first. Change in individual bundle counts can also be calculated. For example, the bundle *if you have a* occurs five times in the first three VBDUs and 13 times in second three VBDUs. (See Appendix A for frequency counts for each bundle.) Calculating change with the formula above, a 160% positive change is shown in the use of this bundle in the second set of units. Alternatively, the bundle *I want to do* occurs 22 times in the first set of VBDUs and eight times in the second set. Here, we see a 64% decline in the use of this bundle as the discourse flows. The frequency and the percent change along with the functional categories were entered in an Excel sheet to create the figure below.

3. Findings

In this section, we first report on the differences in the overall distribution patterns of the bundle functions in the first VBDUs versus the second three VBDUs. Then, we go into details to look at the subtypes of the major functions to see which ones are more robustly apparent in the first set versus the second. Finally, we report on further subcategories and some of the actual bundles that constitute these differences.

3.1 Overall distributional patterns across functional categories

As mentioned above, lexical bundles have been classified into three major categories based on their functions: stance markers, discourse organizers, and referential expressions. In this section, we will discuss the distribution patterns of these three functions between the first three and the second three VBDUs. Table 1 below gives us the raw and percent numbers of this distribution.

Table 1. Frequency distribution of lexical bundle functions and change in percent in the first and the second three VBDUs.

Bundle macro-function	VBDU 1–3	VBDU 4–6	Change %	Direction
stance markers	225	181	19.55	↓
discourse organizers	146	125	14.38	↓
referential expressions	214	261	21.96	↑

As Table 1 above illustrates, overall, about 20% fewer stance expressions and nearly 14% fewer discourse organizers appear in the second set of VBDUs than in the first set. On the other hand, overall, 22% more bundles functioning as referential expressions are apparent in the second three VBDUs than in the first. (See examples for each function and frequency count in Appendix A.)

3.2 Distributional patterns within functional categories

Within the major functional categories, further subcategories were identified. Accordingly, among stance expressions, two further subcategories are identified: epistemic stance expressions and attitudinal stance expressions. Among discourse organizers, we see subcategories such as topic elaboration and directives and topic introduction. Finally, within referential expressions in general, we have categories such as identification/focus, imprecision, specification of attributes, and time/place/text reference (a complete list of categories and corresponding bundles is given in Appendix A).

Figure 1 below illustrates a striking difference in the distribution patterns across the subcategories of function between the first and second sets of units. More specifically, the use of bundles expressing stance generally declines in the second set of VBDUs. Interestingly, epistemic stance bundles such as *I don't know if*, *I don't know what*, and *I think it was* show a nearly 25% decline in the

second set of VBDUs. Other stance bundles expressing attitudes and modality also decline in the second set of VBDUs by about 18%. Examples of these include *what we're going to*, *if you want to*, and *you don't have to*.

In the case of bundles functioning as discourse organizers, we find a 20% drop in the second set of VBDUs for those bundles that introduce topics and focus while there is a 40% growth overall in the bundles that had been associated with a topic elaboration function. For example, the bundle *if you look at* (associated with topic elaboration) appears a total of seven times in the first three VBDUs, and a total of 14 times in the second set of three VBDUs. At the same time, the bundle *I want to do* (associated with topic introduction) appears 22 times in the first three and a total of eight times in the second three VBDUs.

Finally, as Figure 1 also shows, in the case of bundles expressing referential functions, two major subcategories showed a positive change, one declined in frequency as the two sets of VBDUs were compared, and one remained constant. Referential bundles associated with identification and focus show nearly 25% growth, and specification of attributes appear 53% more often in the second set of VBDUs (that is, in the second three units) than in the first set (that is, in the first three units). As an example, bundles such as *was one of the* (associated with identification and focus) appear almost twice as many times in the second set of VBDUs (four and seven times, respectively). Another example is the bundle *a lot of people* (associated with specification of attributes) which appears five times in the first three VBDUs and 15 times in the second three.

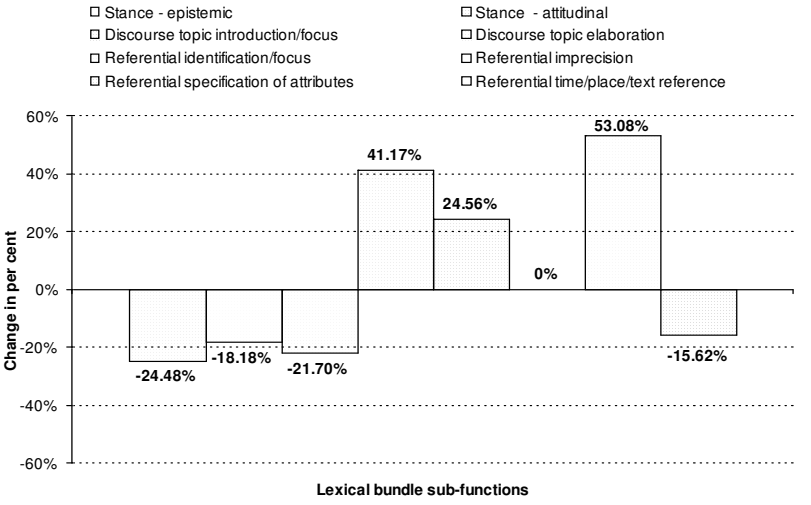


Figure 1. Change in percent between the first and the second three discourse units by lexical bundle subfunctions.

On the other hand, referential bundles associated with time, place, and text reference decline by a little over 15% in the second set of units and those with Imprecision show no change. As an example, the bundle *and things like that*

(associated with imprecision) appear two times in the first set of VBDUs and two times in the second set, respectively. Another example is the bundle *the end of the* (associated with time/place/text reference), which appears 20 times in the first set and eight times in the second set.

3.3 Distributional patterns within declining subcategories overall

We illustrated in the previous section that the use of particular types of bundles increase in the second set of VBDUs while others drop. As seen in Figure 1 above, bundles expressing stance (epistemic and attitudinal/modality); bundles with discourse organizing functions that introduce topic/focus; and bundles expressing time, place, and textual reference are the ones that are used fewer times overall in the second sets of VBDUs. In this short section, we look further into the subcategories of these functions to examine the nature of the bundles. Table 2 below illustrates the general patterns.

Table 2. Frequency distribution and change in percent of overall declining bundle subcategories in the first and second three VBDUs.

Bundle subcategory	**VBDU 1-3**	**VBDU 4-6**	**Change %**	**Direction**
epistemic stance—personal	49	37	24.48	↓
attitudinal/modality stance				
desire—personal	25	16	36	↓
obligation/directive—personal	60	54	10	↓
intention/prediction—personal	29	24	17.24	↓
intention/prediction—impersonal	41	43	4.87	↑
ability—personal	21	7	66.66	↓
topic introduction/focus	129	101	21.70	↓
time/place/text reference				
time	14	17	21	↑
place	8	9	12	↑
multifunctional reference	42	28	33.33	↓

As Table 2 shows, epistemic stance bundles decline in tandem with most of those associated with attitudinal/modality stance bundles. Noticeably, bundles expressing stance with a personal focus, such as personal ability, personal desire, personal obligation/directive, and personal intention/prediction (66.6%, 36%, 10%, 17.24%, respectively), drop dramatically in the second set of VBDUs. In contrast, bundles expressing impersonal intension and prediction show a nearly 5% growth between the two sets of units. If we look at the set of bundles expressing time, place, and text reference, we see that multifunctional references drop by 33.3%, while time and place references grow (21% and 12%,

respectively). Multifunctional bundles are, for example, *the end of the* and *in the middle of*. As listed in Appendix A, an example of place reference is, for example, *in the United States*, and an example of time reference is *at the same time*.

4. Summary and conclusions

This study examined how lexical bundle types are distributed in the first six units of classroom discourse. We relied on existing units of analysis that were presented and characterized in previous literature. That is, we used automatically identified VBDUs as the basis for our discourse segmentation to arrive at patterns of discourse structure, and automatically identified frequently occurring, register-specific word combinations (lexical bundles) as the basis for our distributional pattern. While VBDUs have already been characterized linguistically and functionally via multidimensional analyses, lexical bundles have also been tagged for their most common functions. Accordingly, the first three VBDUs in class discussions have been reported to have very different linguistic features from the next three segments. As Csomay (2005) has shown, the first three VBDUs exhibited interactive discourse and language that reflected contextualized talk with personalized framing features (associated with classroom management functions), while the subsequent units showed monologic talk with language reflecting informational focus and conceptual orientation (associated with the main content focus of the class). Similarly, based on their functions, lexical bundles had been classified into three major functional categories: discourse organizers, stance expressions, and referential expressions (Biber, Conrad, and Cortes, 2004). The novelty in this work is that we tracked lexical bundles with their functions in the first six units of classroom discourse to see the extent to which change in the communicative functions (as identified based on the change in linguistic variation) in discourse structure mentioned above may be supported by lexical patterns as well.

Results showed a difference in lexical bundle use in the two sets of VBDUs. The first set of findings indicates that the use of bundles associated with referential expressions grow steadily in the second set of units, while certain discourse organizers and stance expressions decline. Further, the second set of findings indicates that particular bundles with topic elaboration (discourse organizer) functions and other bundles with specification of attributes (referential) functions nearly doubled in the second set of VBDUs. At the same time, topic introduction (discourse organizer) functions, and time/space/text reference (referential) bundles declined in the second set of VBDUs, as did all stance bundles. Finally, the third set of findings shed light more specifically on those bundles that declined. Interestingly enough, bundles with personal attributes (e.g., personal epistemic stance, personal desire, personal obligation and directive, personal ability, etc.) declined while bundles classified as impersonal intention and prediction showed growth.

The findings of this study are noteworthy from two main perspectives. First, the fact that the patterns of change in the occurrence of the predefined lexical bundle categories of function are in line with earlier linguistic characterizations of the discourse units and their macro-functions is eye opening. That is, lexical bundles not only act as the "building blocks" in discourse on the phrasal and clausal levels (Biber and Conrad 1999:188) but, as they appear with particular frequencies in particular positions, they seem to be prominent participants in following the macro-level change in the discourse structure as well. More specifically, while change in lexico-grammatical patterns between the first set of three VBDUs and the second set (see above) in classroom talk clearly corresponded to change in discourse functions as reported in earlier work (Csomay, 2005, 2007), this pattern of variation and the corresponding change in focus in the discourse seems to be marked by the difference in the most frequently used lexical bundle types between the two sets of VBDUs as well. As the results of this study show, bundles associated with elaboration and clarification, as well as those associated with specificity, increase in the second set of units, while the number of bundles associated with personal stance decline. This pattern of change in the frequency of lexical bundle types between the two VBDU sets stands in direct relation with earlier findings on the relationship between intra-textual linguistic variation and discourse functions (Csomay, 2002, 2005, 2007; Biber, Connor, Jones, and Upton, 2007).

Related to this, secondly, the fact that functions of empirically identified lexical patterns support other lexico-grammatical patterns and their communicative functions in discourse structure is indeed a promising ground for further lexical studies. More specifically, the findings of this study call for further investigations of the relationship between the positions of particular lexical items in the discourse structure and their discourse functions. The empirical question of whether we are able to identify discourse functions relying on the frequency and positioning of lexical items remains to be explored further.

5. Notes

1. "Conceptually, a *Vocabulary-Based Discourse Unit* (VBDU) is a block of discourse defined by its reliance on a particular set of words. The boundary of a VBDU is identified as the place in a text where the author/speaker switches to a new set of words. Because the topic of discourse is expressed through vocabulary, VBDUs can usually be interpreted as topically-coherent units" (Biber, Connor, Csomay, Jones, Keck, and Upton, 2007: 156).

References

Biber, D. (2006), *University language: a corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, D., and S. Conrad (1999), "Lexical bundles in conversation and academic prose," in: H. Hasselgard and S. Oksefjell (eds.) *Out of corpora*. Amsterdam: Rodopi, 181–190.

Biber, D., S. Conrad, and V. Cortes (2004), "'If you look at ...': lexical bundles in university teaching and textbooks," *Applied linguistics*, 25: 371–405.

Biber, D., U. Connor, and T. Upton (2007), *Discourse on the move*. Amsterdam: John Benjamins.

Biber, D., U. Connor, E. Csomay, J. K. Jones, C. Keck, and T. Upton (2007), "Introduction to the identification and analysis of vocabulary-based discourse units," in: D. Biber et al. (eds.) *Discourse on the move*. Amsterdam: John Benjamins, 155-173.

Biber, D., E. Csomay, J. K. Jones, and C. Keck (2004), "A corpus linguistic investigation of vocabulary-based discourse units in university registers," in: U. Connor and T. Upton (eds.) *Applied corpus linguistics: a multidimensional perspective*. Amsterdam: Rodopi, 53–72.

Biber, D., S. Conrad, R. Reppen, P. Byrd, and M. Helt (2002), "Speaking and writing in the university: a multi-dimensional comparison," *TESOL quarterly*, 36: 9–48.

Biber, D., S. Johansson, G. Leech, S. Conrad, and E. Finegan (1999), *The Longman grammar of spoken and written English*. London: Longman.

Camiciottoli, C. B. (2004), "Walking on unfamiliar ground: interactive discourse markers in guest lectures," in: A. Partington et al. (eds.) *Corpora and discourse*. Bern: Peter Lang, 91–106.

Cortes, V. (2004), "Lexical bundles in published and student disciplinary writing: examples from history and biology," *English for specific purposes,* 23: 397–423.

Cortes, V., and E. Csomay (2007), "Positioning lexical bundles in university lectures," in: M. C. Campoy and M.J. Luzon (eds.) *Spoken corpora in applied linguistics.* Bern: Peter Lang, 57–76.

Csomay, E. (2002), "Episodes in university classrooms: a corpus-linguistic investigation", Ph.D. dissertation, Northern Arizona University.

Csomay, E. (2005), "Linguistic variation within university classroom talk: a corpus-based perspective," *Linguistics and education*, 15: 243–274.

Csomay, E. (2006), "Academic talk in American university classrooms: crossing the boundaries of oral literate discourse?," *Journal of English for academic purposes*, 5: 117–135.

Csomay, E. (2007), "Vocabulary-based discourse units in university class sessions," in: D. Biber et al. (eds.) *Discourse on the move*. Amsterdam: John Benjamins, 213–238.

Fortanet, I. (2004), "The use of 'we' in university lectures: reference and function," *English for specific purposes*, 23: 45–66.

Hearst, M. (1997), "TextTiling: segmenting text into multi-paragraph subtopic passages," *Computational linguistics*, 23: 33–64.

Mauranen, A. (2001), "Reflexive academic talk: observations from MICASE," in: R. Simpson and J. Swales (eds.) *Corpus linguistics in North America: selection from the 1999 symposium*. Ann Arbor: University of Michigan Press, 165–178.

Nesi, H., and H. Basturkmen (2006), "Lexical bundles and discourse signalling in academic lectures," *International journal of corpus linguistics*, 11: 147–168.

Prince, E. (1981), "Toward a taxonomy of given-new information," in: P. Cole (ed.) *Radical pragmatics*. New York: Academic Press, 223–56.

Simpson, R., and D. Mendis (2003), "A corpus-based study of idioms in academic speech," *TESOL quarterly*, 37: 419–441.

Swales, J., and B. Malczewski (2001), "Discourse management and new-episode flags in MICASE," in: R. Simpson and J. Swales (eds.) *Corpus linguistics in North America: selection from the 1999 symposium*. Ann Arbor: University of Michigan Press, 145–164.

Swales, J., and A. Burke (2003), "'It's really fascinating work': differences in evaluative adjectives across academic registers," in: P. Leistyna and C. Meyer (eds.) *Corpus analysis: language structure and language use*. New York: Rodopi, 1–18.

Youmans, G. (1991), "A new tool for discourse analysis: the Vocabulary Management Profile," *Language*, 67: 763–789.

Appendix

Lexical bundles identified in the first six Vocabulary-Based Discourse Units (VBDUs) grouped according to their functions

Bundle function	**VBDU 1 to 3**	**VBDU 4 to 6**
1. Stance markers		
A. Epistemic stance		
Personal		
*and I think that***	3	3
*I think it was***	6	3
*you know what I***	4	4
*I don't know if***	12	5
*I don't know what***	8	5
*I don't know how***	5	4
*I don't know I***	4	3
*and I don't know***	7	10
Total epistemic stance	**49**	**37**

B. Attitudinal/modality stance		
B1. Desire—personal		
*if you want to***	20	14
*I don't want to***	5	2
B2. Obligation/directive—personal		
*going to have to***	8	2
*and you have to***	3	4
*you have to do***	2	3
*you look at the***	6	5
*you might want to***	3	4
*I want you to****	10	14
*you have to be***	6	5
*you need to know***	2	1
*you don't have to***	16	10
*you don't want to***	4	6
B3a. Intention/prediction—personal		
*I'm not going to***	0	0
*we're going to do***	9	9
*we're going to have***	5	3
*and we're going to***	4	5
*what we're going to***	11	7
B3b. Intention/prediction—impersonal		
*is going to be****	9	19
*not going to be***	8	6
*going to be a***	8	4
*are going to be***	8	3
*going to be the***	2	8
*it's going to be***	6	3
B4 Ability—personal		
*to be able to****	14	4
*to come up with***	7	3
Total attitudinal/modality stance	**176**	**144**
2. Discourse organizers		
2A. Topic introduction/focus		
*I want to do***	22	8
*going to talk about***	15	8
*want to do is***	14	6
*what I want to***	14	10
*take a look at***	14	5
*what do you think***	11	4

*a little bit about***	10	12
*if we look at***	3	6
*if you look at***	7	14
*if you have a***	5	13
*to go ahead and***	4	0
*to look at the***	2	4
*you know if you***	3	5
*want to talk about***	5	6
Total topic introduction/focus	**129**	**101**
2B. Topic elaboration/clarification		
*on the other hand***	4	10
*has to do with***	3	3
*you know I mean***	5	1
*to do with the***	1	5
*I mean you know***	4	5
Total topic elaboration/clarification	**17**	**24**
3. Referential expressions		
3A. Identification/focus		
*one of the things****	16	15
*those of you who***	10	6
*of the things that****	10	15
*and this is a***	3	4
*is one of the***	6	7
*was one of the***	4	7
*and this is the***	3	7
*and one of the***	4	4
*that's one of the***	1	6
Total referential identification and focus	**57**	**71**
3B. Imprecision		
*or something like that***	4	8
*and stuff like that***	6	2
*and things like that***	2	2
Total referential imprecision	**12**	**12**
3C. Specification of attributes		
3Ca. Quantity		
*a little bit of***	16	9
*a lot of people***	5	15

*have a lot of***	3	10
*and a lot of***	4	8
*a little bit more***	11	10
*a lot of the***	5	11
*how many of you***	6	12
*the rest of the***	5	9
*greater than or equal***	4	4
*than or equal to***	4	7
*in a lot of***	1	4
*there's a lot of***	8	7
*a lot of times***	0	3
3Cb. Intangible framing attributes		
*in the case of***	5	6
*in terms of the***	4	9
Total specification of attributes	**81**	**124**
D. Time/place/text reference		
D1. Time		
*at the same time***	14	17
D2. Place		
*in the united states***	8	9
D3. Multifunctional reference		
*the end of the***	20	8
*at the end of***	17	14
*in the middle of***	5	6
Total time/place/text reference	**64**	**54**

From the functional classification of common lexical bundles in university lectures (adapted from Biber, Conrad, and Cortes, 2004)

Key to symbols:
** = 40–99 per million words
*** = over 100 per million words

Suggestions and recommendations in academic speech

Luciana Diniz

Portland Community College

Abstract

This chapter investigates the communicative functions of modals (e.g., should*,* could*), as well as lexical verbs, such as* recommend *and* suggest*, as they are used by faculty in academic spoken discourse. Results show that these expressions are frequently used by professors to communicate indirect orders (as opposed to directives) in a number of situations. Several speech events from the* Michigan Corpus of Academic Spoken English *(MICASE) were selected to constitute the corpus, including academic lectures, office hours, dissertation defenses, and seminars. Wordsmith Tools 5.0 (Scott, 1999) and the search tools from the MICASE web site were used to identify phrases and semifixed phrases containing the target modals and lexical verbs. The results have immediate pedagogical applications to both English for Academic Purposes (EAP) and International Teaching Assistant (ITA) education.*

1. Introduction

Research has shown that modal verbs can acquire a variety of meanings depending on the context in which they are encountered (e.g., Hinkel, 1995; Thomson, 2000). This wide range of functions imposes a challenge for nonnative English speakers (NNES), who not only have to decode the linguistic portion of discourse, but also the social expectations of language users in order to understand the nuances of the message being communicated to them. In fact, even NNES with an advanced level of proficiency in their second language still have difficulties with the complexity of modal meanings, which are often context- and culture-specific (Hinkel, 1995).

Biber et al. (1999) group modals in two categories: central modals and semi-modals. Central modals include *can, could, may, might, shall, should, will, would,* and *must,* while semi-modals consist of phrases that function like modals (e.g., *need to, have to, be supposed to*). As to their meanings, modals fall into three major groups: (1) permission/possibility/ability (e.g., *can, could, might*); (2) obligation/necessity (e.g., *must, should, need to, ought to, be supposed to*); and (3) volition/prediction (e.g., *will, would, shall, be going to*). However, Biber et al. explain that in language use, the meanings of modals are much more complex and go beyond these fundamental categories.

In an effort to simplify the functions of modals and make them more accessible to language learners, widely used pedagogical grammar books and textbooks such as Azar (2000); Celce-Murcia and Larsen-Freeman (1998); Fragiadakis, Rosenfield, and Teimroth-Zavala (2004); Fuchs and Bonner (2005);

and Raimes (2004) attempt to group modals into clear-cut categories. For example, *may, might,* and *could* are often labeled as possibility, expressing a weak degree of certainty. *Should, ought to,* and *had better,* on the other hand, indicate *advisability,* while *can* indicates *ability*.

In such texts, the modals that indicate advisability are frequently used in contexts that describe suggestions which may or may not be accepted by the listener. These modals are frequently contrasted with *must* and *have to,* which encompass a stronger sense of necessity and obligation. The majority of the examples containing modals of advisability provided in these texts show that the acceptance of the suggestion is deemed as optional and do not necessarily imply any serious consequences to the listener.

Another way language users can express advisability is through lexical verbs such as *suggest* and *recommend*. These verbs are often combined with a modal to convey the same idea (e.g., "I would definitely recommend that restaurant near my house."). The *MacMillan English Dictionary* (Rundell, 2004) defines *suggest* as "to tell someone about something that may be useful or appropriate for a particular purpose" (e.g., "Can you suggest an inexpensive restaurant?") and *recommend* as "to advise someone that they should do something" (e.g., "I recommend that you buy a more powerful computer."). These definitions and examples demonstrate that the most usual meanings of *suggest* and *recommend* express a similar idea as the modals of advisability presented in pedagogical grammars.

While it is true that these modals and lexical verbs most often denote advisability, these meanings do not account for all their functions in specific contexts. For one, these explanations and definitions do not take into consideration the social roles of the language users. Celce-Murcia and Larsen-Freeman (1998) point out that most modals are deeply influenced by social functions and require, therefore, that language users take into account the social situation in order to use them properly. In an interaction between a teacher and a student in the academic setting, for instance, suggestions and recommendations may acquire a stronger degree of advisability owing to the discourse role performed by the teacher. In this context, suggestions are often interpreted as orders because of the power status involved in the dynamic between instructors and students. If students do not follow the "recommendations" made by the teacher, they are less likely to succeed in that particular assignment or class. In other words, in the case of teacher-student interactions, utterances that may be interpreted as mere suggestions and recommendations elsewhere are understood as orders in the academic setting, showing how context-specific the meanings of modals and lexical verbs can be.

2. Indirect orders

Teachers use both direct and indirect language to communicate their expectations in the classroom. Some examples of direct language include imperatives (e.g.,

"Open your books.") and modals such as *must* and *have to,* which often convey a strong meaning of obligation. In a study of teachers' use of direct language, Reppen (2008) found a great number of directives not only in instructors' speech but also in syllabi and teacher-made handouts.

Many times, however, teachers choose to employ more subtle language to communicate their expectations. For instance, they use less face-threatening constructions such as *you might* or *you should* rather than directives such as *you must, you have to,* or *I want you to,* or even a blend of indirect language; see example (1) to express indirect orders. Examples (1), (2), (3), and (4) are from the *Michigan Corpus of Academic Spoken English* (MICASE). The information in italics indicates the speech event from which the example was taken.

(1) I might wanna suggest … (Statistics Office Hour)
(2) I recommend you use these … (Biopsychology Lab)
(3) So I suggest you to do, to read, to print out, the solution for this chapter … (Economics Discussion Section)
(4) … you may and might include that in your report … (Biology of Birds Discussion)

In Example (2), the teacher is recommending the use of a certain tool to perform the lab experiment, while in Example (3), the teacher is suggesting that students pay special attention to the solutions for the exercises in that particular chapter. In Example (4), the teacher is communicating his/her expectations toward a particular assignment. All these situations can be interpreted more strongly than a simple suggestion or recommendation. In order to succeed in that specific lab experiment, the students will most likely have to use that particular tool, while studying the exercises from that specific chapter and including the precise information in that report will most likely help the students to do well on the exam or assignment or in the class.

The objective of this chapter is to identify fixed and semi-fixed phrases that instructors use to communicate indirect orders as well as to analyze the contexts in which these phrases appear in order to gain a deeper understanding of the reasoning behind teachers' language choice.

3. Methodology

3.1 The corpus

Seven speech events from MICASE—dissertation defenses, discussions, lab sections, large lectures, small lectures, office hours, and seminars—constitute the corpus used in this study. These speech events were selected because they represent clear cases of interactions between professors and students. The total number of words was approximately 1,100,000 (see Table 1).

Table 1. Speech events (from MICASE) that constituted the corpus.

Speech Event	Number of Words
dissertation defenses	56,837
discussions	74,904
lab sections	73,815
large lectures	251,632
small lectures	333,338
office hours	171,188
seminars	138,626
Total	1,100,340

3.2 Lexical verbs and modals

In order to identify the phrases and semi-fixed phrases functioning as indirect orders, a combination of a corpus-based and a corpus-driven approach was used (Tognini-Bonneli 2001). A corpus-based approach uses the corpus as an instrument to investigate and give support to (or refute) assumptions and intuitive ideas from the researcher, while a corpus-driven approach allows the results to emerge from the corpus without preconceived categories. Since it was not feasible to begin the search without having some preliminary categories, a combination of these two approaches was found suitable for the objectives of this paper. This entailed the following steps:

- A list containing words and phrases included in previous research and personal teaching experience was created (see Appendix A).
- A list of advice areas that could possibly contain phrases indicating indirect orders was generated (see Appendix B). Concordance lines (and extended context) containing these advice areas were analyzed, and phrases preceding these words that functioned as indirect orders were added to a list.
- Samples of randomly selected texts in the corpus were read and a list of target words/phrases was compiled.
- N-grams and semi-fixed phrases containing the target words and phrases were detected using Wordsmith Tools 5.0 (Scott, 1999).

Following the compilation of all words and phrases that potentially indicated indirect orders, a close analysis of their concordance lines (as well as their expanded context) was performed. A careful and closer analysis of the extended context in each instance was necessary in order to determine whether a particular phrase was really performing an indirect order function. Only utterances that presented clear evidence of communicating indirect orders were considered.

4. Results

As predicted, the identified words and phrases fell into two categories—namely, modal verbs and lexical verbs—and the results are shown in Table 2. The numbers in parentheses indicate the total number of instances with which each specific word occurred in the corpus (raw frequency).

Table 2. Identified words and phrases in the MICASE subcorpus.

Modals	Lexical verbs
might/might not (1,046)	*suggest* (66)
should/not/shouldn't (1,200)	*recommend* (19)
may/may not (820)	*encourage* (13)
could/not/couldn't (2,089)	*consider* (96)
can/not/can't (6,525)	*expect* (141)
would/'d/not/wouldn't (6,909)	*hope* (127)
	like (7,481)
	prefer (33)
	want (2,594)
	urge (4)

Table 3 shows the most frequently identified lexical verbs ordered by percentage of occurrences—that is, the frequency in which the lexical verb was encountered in a context that communicated indirect order. This ordering takes into consideration the total number of instances of that specific word in the corpus. The table also presents the most common phrases containing the lexical verbs when conveying indirect orders.

Table 3. Identified lexical verbs ordered by percentage of occurrence in the MICASE subcorpus.

Verb	Frequency	Indirect order	Most common phrases
encourage	13	7 (~53%)	*I would encourage* (3) *I want to encourage* (2)
recommend	19	5 (~26%)	*I (would, really) recommend* (3)
suggest	66	17 (~25%)	*I would suggest* (8)
urge	4	1 (~25%)	*I urge you to* (1)
expect	141	12 (~8.5%)	*I expect you to* (6)
hope	127	5 (3.9%)	*I would hope* (2)
prefer	33	2 (~6%)	*I would prefer* (2)
consider	96	4 (~4%)	*I want (you) to consider* (4)
want/wanna	2,594	74 (~3%)	*you might wanna/want to* (24)
like	7,481	25 (0.3%)	*I would like you to* (22)

The lexical verbs described in Table 3 co-occur with modal verbs in approximately 50% of the instances, forming fixed and semi-fixed phrases that communicate indirect orders. Some lexical verbs tend to acquire the meaning of

indirect order more often when combined with a modal verb. For example, all 25 instances of the lexical verb *like* that communicated indirect order, were preceded by the modal *would* (or the abbreviated form *'d*).

In Examples (5), (6), and (7), the modals *would* and *may*, when combined with the lexical verbs *recommend*, *prefer*, and *consider*, soften the message of obligation even more. In Examples (5) and (6), the instructor is communicating his/her expectations for an assignment. In Example (7), the teacher is asking the student to eliminate an idea from an essay.

(5) … well, I would recommend that you choose one and not, not do both, cuz that way, if you, do do both you're introducing an uncontrolled variable. (Linguistics Independent Study)

(6) … we impute, values for that. and I- I'd prefer, that you don't just impute the grand mean because if it's a Catholic school the grand mean is not really relevant… (Statistics in Social Science Lecture)

(7) … and you may wanna consider removing that … (Artificial Intelligence Dissertation Defense)

Unlike the target lexical verbs, not all concordance lines containing modals were analyzed, given the large number of instances of modals in the corpus. In order to select relevant instances of indirect orders including modals, the procedure outlined in Section 3.2 was followed.

5. Functions of indirect orders

A close analysis of the selected concordance lines revealed that the utterances containing indirect orders performed a number of communicative functions, including *gaining time, being polite, facilitating discussion, highlighting information, transferring responsibility,* and *communicating expectations*. In almost all cases, the same utterance performed more than one function at the same time. However, in order to simplify their descriptions, the functions were broken down into separate categories. Table 4 shows the relative frequencies of each of the functions in utterances containing the target lexical verbs.

Table 4. Functions of indirect orders (containing the target lexical verbs, optionally preceded by modals) in the MICASE subcorpus

Function	Frequency
being polite	103
transferring responsibility	83
communicating expectations	46
highlighting information	32
gaining time	24
facilitating discussion	4

Relative frequencies of phrases containing modals (without the target lexical verbs) are not provided because a different quantitative approach was used to detect such utterances (see Sections 3.4 and 4).

5.1 Gaining time

Many phrases expressing indirect order in the subcorpus are surrounded by vague language, such as *sort of* and *kind of*, as well as repetitions. Among other reasons, teachers seem to favor the use of such constructions in order to gain time to process the information they will say next. In Examples (8) and (9), the teacher is having one-on-one conversations with students during an office hour and is trying to show them that their papers need a stronger introduction/conclusion. The phrases containing vague language are underlined.

(8) … you might just wanna like again sort of draw out some conclusions from this a little bit … (Anthropology of American Cities Office Hours)

(9) … so that is the beginning of your introduction and that's cool. Um, you might wanna sort of expand it a little bit thinking about these sort of themes… (Anthropology of American Cities Office Hours)

In Example (10), the teacher is communicating the kind of information he/she expects a specific paper to contain. It could be argued that the vague expression *I'd kinda like*, as well as the other phrases in Examples (8) and (9), is not only used to gain time, but also as a politeness strategy, softening the utterance and not letting the instructor come across as rude when giving the command. This strategy is discussed in the next section.

(10) I'd kinda like to hear what your measures of its success are … (Graduate Public Policy Seminar)

5.2 Being polite

Because modals and lexical verbs indicating advisability are usually less face threatening than those designating strong obligation, this type of language is often used as a politeness strategy. Many times, teachers choose indirect language over directives in an apparent attempt to save face when communicating to students that their work is not good enough. For instance, in Example (11), the teacher is indicating the need of improvement in the organization and content of the essay paragraph. It seems that the instructor uses indirect language, even if subconsciously, to avoid discouraging the student, making sure she/he emphasizes both the positive aspects of the paper and the changes to be made.

(11) I wouldn't rewrite the whole paper, but maybe I would would take, um I would take th- this first page and the first paragraph on the second page, and, and I'd begin by saying, what do_ what's the job I want that paragraph to do? (Intro to Poetry Office Hours)

In an apparent effort to save face, instructors tend to use hedging devices such as *I think you might/should* or *you might think about.* The indirect orders with additional hedging devices are probably chosen to avoid embarrassment and to establish rapport with the students. For instance, in Examples (12) and (13), the instructor uses a number of mitigation devices such as *I think* and *may + wanna* rather than simply *may consider* to indicate changes that have to be made.

(12) I think you might give the example more precisely … (History Review Discussion Section)
(13) … you may wanna consider removing that … (Artificial Intelligence Dissertation Defense)

In Examples (14) and (15), the instructor utilizes a series of repetitions, probably searching for constructions that are less threatening. It could be argued that the teacher is also trying to *gain time* in these examples, which gives further evidence to the multifunctionality of utterances explained in Section 5.

(14) … um, but y- y- you might want to expand that a little bit because I think you you lean towards it but you don't, re- de- dig, deep into it. (Music Dissertation Defense)
(15) … so, uh, I suppose I would do some work pulling this part togeth- uh, anyway this part together [Student: uhuh] into an argument about sort of how it's representing the community first. And then I'd go back … (Anthropology of American Cities Office Hours)

The n-gram *I would suggest* (seven instances in the corpus) is quite frequent in this type of function, as shown in Examples (16) and (17). In Example (16), the instructor is communicating his expectations regarding an upcoming assignment. In Example (17), the teacher is asking students to take notes, probably because these notes will be part of a future graded task.

(16) … in this first one so I would suggest, no more than maybe, three or four child-level variables and, three to four, uh school-level variables … (Statistics in Social Sciences Lecture)
(17) … I think that's satisfactory, so I would, suggest taking notes as you go along so you can remember um, what kind of questions or issues the exhibits… (Archeology of Modern American Life Lecture)

5.3 Facilitating discussion

Indirect orders are also used to facilitate discussions between teachers and students. This less intimidating type of language seems to give some balance to the power relationship and open some space for students to interrupt to make comments, ask questions, and express their opinions. In Example (18), the teacher uses a variety of indirect language. Given the number of interruptions, it appears

that the student feels quite comfortable discussing the improvements on his/her essay with the instructor.

(18) Instructor: or or you could say unlike older residents in Howell, who depend on groups, community groups and events, residents of Novi go to the mall to socialize. That would work. You just gotta make, you gotta put H8owell in there [Student: okay] instead of just people. Um, good. Uh going to the mall is like, is, is an event [Student: in itself yeah] in and of itself. Um more (specific) uh ... uh juh juh juh juh juh um, [Student: I couldn't figure out how to say that] is an event in itself, a social event.
Student: Okay. Alright that'll work. (Anthropology of American Cities Office Hours)

In Example (19), the indirect language is used by both the student and the teacher. The student says, "What would you recommend?" (rather than "Can you tell me what to do?") to ask for the instructor's guidance in a subtle and polite manner.

(19) Student: so what sh- what would you recommend?
Instructor: well I would look at the exam two that's in the back [Student: okay.] of your book a- an- of the packet.
Student: Of that model? (Statistics Office Hours)

5.4 Highlighting important information

In many cases, indirect orders are also used to emphasize important information that is likely to help students succeed in the class. In Example (20), the teacher points out a common confusion of the experiments in the book:

(20) I urge you to make a note in your notes not to confuse experiments on page twenty-one and twenty-six … (Biology of Cancer Lecture)

When used to highlight important information, indirect orders are sometimes preceded by intensifiers such as *strongly*, *highly,* or *really* (six instances in the corpus), as in Examples (21) and (22). A close examination of the sound files that accompany MICASE revealed that instructors tend not to stress the intensifiers in their speech, making it more difficult for students to rely on this paralinguistic cue. None of the six instances of intensifiers was stressed in the corpus.

(21) … so I strongly recommend, you leave a piece of area in your notes so that we can fill this in … (Biology and Ecology of Fishes Lecture)

(22) … we do highly suggest that you write them up. For class it enables you to prepare. You then have a record of the case, for your notes for later … (Behavior Theory Management Lecture)

5.5 Transferring responsibility

Instructors also use indirect orders as a way of transferring some of the learning responsibility to the students. By using phrases with the pronoun *you*, including *you might want to* (8 instances in the corpus) and *you might wanna* (16 instances), teachers are sharing the responsibility of the assignment. In Examples (23) and (24), it is possible to notice that, when choosing expressions such as *you'll wanna write* and *you wanna somehow*, instructors seem to establish that the students are the ones responsible for improving their papers.

(23) …then you'll wanna write something about what you've learned […] you'll wanna have a cover page […] when you write up your paper, you'll want to lay out your research question … (Statistics in Social Science Lecture)

(24) This is a tiny bit repetitive in here, so you might just wanna somehow tighten it up … (Anthropology of American Cities Office Hours)

5.6 Communicating expectations

Another, sometimes overlapping, function of indirect language is the communication of the instructor's expectations of a certain aspect of the class. In Examples (25) and (26), *I would hope to hear* (one instance in the corpus) and *I would like to see* (seven instances in the corpus) are utilized to indicate the exact information expected in the assignments.

(25) … establishing cause and effect, uh so I would hope to hear a variety of approaches, which you might come up with, in terms of trying to experimentally address that question… (Biology of Cancer Lecture)

(26) … how do you calculate these local losses, these friction losses? What coefficients do you use? Right. I would like to see all that. (Hydraulics Problem Solving Lab)

6. Summary and pedagogical implications

As the results demonstrate, the use of indirect language to give orders is quite common in an academic setting. Modals and lexical verbs that are usually labeled as performing a simple *advisability* function in pedagogical grammars are, in reality, context-sensitive. They acquire stronger meanings of obligation in many contexts in academia when uttered by an authoritative figure such as the teacher. These meanings are commonly accepted and shared in the academic language community.

The results also show that teachers probably use indirect language as an attempt to reduce the power differential between them and the students. The less threatening constructions seem to function as hedging mechanisms for teachers to

communicate their expectations while also demonstrating respect for students' individuality.

Perhaps the use of these constructions is a consequence of recent tendencies in education in which teachers are inclined to change their role into facilitators rather than rule dictators. However, even when playing the facilitator role and giving the students more freedom to express their opinions and ideas, teachers are still the ones responsible for grading them. This way, the instructor is still the authoritative figure whose orders (direct and indirect) should be followed. Students, therefore, have to be aware of this power relationship that permeates the classroom dynamics in order to effectively interpret the hidden messages conveyed by the teacher when it comes to advisability. If students fail to recognize these orders, the resulting misunderstanding can be frustrating for both teachers and students. Instructors, realizing that their expectations were not met (and being confident that they expressed them very clearly), may perceive students who do not comply with their orders as slackers and penalize them with a bad grade.

Given the evident intricacy of meanings of advisability modals and lexical verbs in the academic setting, it is not surprising that their uses are challenging to NNES, as shown by Hinkel (1995). This complexity is reinforced by pedagogical grammars and textbooks presentations of such verbs, which tend to label modals of advisability as simple suggestions, not taking into account the context or the role of the participants in the discourse.

The constructions described in this article are not always used as indirect orders, which may also contribute to the challenging aspect of mastering the meaning of modals. For instance, in Examples (27) and (28), instructors are indeed suggesting books for students to read in case they are interested in the particular subject discussed in the class. Not following the teachers' advice in these particular cases will most likely not have any impact on students' grades.

(27) … uh in that case <u>let me recommend</u> a book called *Bridge of Birds.* It's a fantasy novel. I use it in my, uh, metaphors class … (Linguistics Independent Study)

(28) … um and then <u>I wanted to give you</u> on the back <u>some suggestions</u> for additional reading. May or may not help you in your, um, research projects because … (Archeology of Modern American Life Lecture)

A more controlled study in which second language learners would have to rate the degree of advisability of selected teacher/students interactions would be necessary in order to confirm whether NNES indeed have difficulties understanding the meanings of indirect orders. However, as Hinkel (1995) pointed out, addressing culture, values, participants' roles, and social expectations when teaching modal meanings can certainly facilitate students' understanding of this complex lexicogrammatical topic. Contrasting conversations between friends, teacher-student, coworkers, doctor-patient, and employee-employer, among

others, might be helpful for students to understand the power relationships established in different discourse communities.

7. Notes

1. International teaching assistants (ITAs) may also benefit from learning the functions of indirect language described here. ITAs can be presented with politeness strategies, which can be utilized for more efficient communication with their own students.

References

Azar, B. (2000), *Understanding and using English grammar*. New Jersey: Prentice Hall.

Biber, D., S. Johansson, G. Leech, S. Conrad, and E. Finegan (1999), *Longman grammar of spoken and written English*. Essex, England: Pearson Education Limited.

Celce-Murcia, M., and D. Larsen-Freeman (1998), *The grammar book*. New York: Heinle and Heinle.

Fuchs, M., and M. Bonner (2005), *Focus on grammar*. New York: Pearson ESL.

Hinkel, E. (1995), "The use of modal verbs as a reflection of cultural values," *TESOL quarterly*, 29(2): 325–343.

Kalkstein, H., E. R. Fragiadakis, and C. Graham (2004), *Grammar step by step 3*. New York: McGraw-Hill.

Raimes, A. (2004), *Grammar troublespots: a guide for student writers*. Cambridge: Cambridge University Press.

Reppen, R. (2008), "'Students must': a corpus based look at directives in university language," Paper presented at the 2008 meeting of the AACL, Provo, UT.

Scott, M. (1999), *Wordsmith Tools 5.0*. Oxford: Oxford University Press.

Simpson, R. C., S. L. Briggs, J. Ovens, and J. M. Swales (2002), *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI: The Regents of the University of Michigan.

Rundell, M. (ed.) (2004), *Macmillan English dictionary for advanced learners of American English*. London: Palgrave Macmillan.

Thompson, P. (2000), "Modal verbs in academic writing," in: B. Ketteman and G. Marko (ed.) *Teaching and learning by doing corpus analysis.* Amsterdam: Rodopi, 305–325.

Tognini-Bonelli, E. (2001), *Corpus linguistics at work*. Amsterdam: John Benjamins.

Appendix

Appendix A. List of advice areas.

assignment, *check*, *final paper*, *final project*, *homework*, *include*, *on time*, *read*, *study*, *write*

Appendix B. List of Previews Research and Personal Teaching Experience expressions.

I suggest that you, I think you should, I think you should, I would, I'd like you to, if I were you, you may want to think about, you might want to, you should, you should consider

Building a forensic corpus to test language-based indicators of deception

Eileen Fitzpatrick

Montclair State University

Joan Bachenko

Deception Discovery Technologies

Abstract

Experimental laboratory results, often performed with college student subjects, have proposed several linguistic phenomena as indicative of speaker deception. We have identified a subset of these phenomena that can be formalized as a linguistic model. The model incorporates three classes of language-based deception cues: (1) linguistic devices used to avoid making a direct statement of fact, for example, hedges; (2) preference for negative expressions in word choice, syntactic structure, and semantics; (3) inconsistencies with respect to verb and noun forms, for example, verb tense changes. The question our research addresses is whether the cues we have adapted from laboratory studies will recognize deception in real-world statements by suspects and witnesses.

The issue addressed here is how to test the accuracy of these linguistic cues with respect to identifying deception. To perform the test, we assembled a corpus of criminal statements, police interrogations, and civil testimony that we annotated in two distinct ways, first for language-based deception cues and second for verification of the claims made in the narrative data. The paper discusses the possible methods for building a corpus to test the deception cue hypothesis, the linguistic phenomena associated with deception, and the issues involved in assembling a forensic corpus.

1. Introduction

The ability to detect deceptive statements in text and speech has broad applications in law enforcement, intelligence gathering, and human resources. Recent work, primarily in psychology and criminal justice, has indicated that detecting this deception is possible because liars "leak" cues to their deceit through facial and body movements, vocal changes, and verbal choices (see DePaulo et al., 2003, for a review). Among the verbal choices are discrete attributes such as hedging, tense changes, and changes in referring expressions; at the global level, factors such as coherence, narrative length, and narrative balance have also been shown to be operative.

Currently, approaches to language-based deception detection come from two communities: experimental psychology and law enforcement. Studies in experimental psychology largely focus on laboratory experiments with subjects,

usually students, acting out a scenario. Because the basic facts of the scenario are controlled, it is possible to observe subjects' linguistic behavior in a uniform test and so establish statistically the language features that are most likely to be associated with deception. However, for ethical reasons, laboratory experiments lack high-stakes pressures: the subjects have nothing to lose if they are caught lying.

Many in law enforcement use language-based methods—grouped under the heading of *statement analysis*—to discover deception in an interview or narrative (Adams, 2002; Smith, 2001). In this case, possible liars have a great deal to lose if they are caught: reputation, money, freedom, job. However, ground truth is often an unknown unless the facts emerge over time through continued investigation, suspect cooperation, or luck. This lack of control over basic facts makes it nearly impossible to conduct a controlled experiment using real-world data. As a result there are few scientific studies of statement analysis as it is used in real-world applications.

Our analysis incorporates aspects of both approaches. In Bachenko et al. (2008) we describe a model of deceptive language that was tested against ground truth in a corpus of criminal and civil narratives. The core assumption of this work is that experiments on corpora can yield probabilistic models that serve as predictors of linguistic behavior. For this approach to succeed, however, the corpora must be large enough to produce credible results.

We consider here whether a model built from data involving verbal choices that includes a combination of a broad array of cues can predict with a high degree of probability whether a given proposition is truthful or deceptive. Our primary focus is on the type of test that would support the design of such a model and the data needed for the test.

Obtaining data that includes such cues is difficult enough, but deception data that enables the testing of correlations between cues and deception has to face a more difficult hurdle: it has to verify the truth or falsity of the statements, the so-called "ground truth." This paper describes the building of a corpus of potentially deceptive statements from real-world data—criminal statements, police interrogations, depositions, and congressional testimony—that is then annotated twice: for deception cues and for ground truth.

The difficulty of assembling ground truth data goes beyond standard linguistic methods to the investigation of the facts behind the narrative. The difficulty persists even when lying is limited to deliberate attempts to mislead since the experimenter must be able to attest to the external evidence that supports the true/false judgment.

The problems of obtaining ground truth have led psychologists to devise experiments that enable control and observation of the facts behind a subject's statements. However, experimental controls tend to remove the data gathered from real-world situations that simulate the high-stakes scenarios that the legal and intelligence applications demand.

We consider a typology of deception experiments first presented by Miller and Stiff (1993) in order to home in on a design that would enable the gathering

of both verbal data and ground truth data in a real-world setting that provides the high stakes needed to motivate the subject to lie.

We discuss the design chosen, the gathering of the data, the establishment of ground truth for the propositions in the data, and the building and testing of the probabilistic model to predict deception.

2. Protocols for gathering deceptive data

The typology of scenarios used to test deception, according to Miller and Stiff, ranges from the highly controlled to the essentially free form. We exclude from consideration scenarios where ground truth testing is not possible, namely, scenarios in which subjects express personal feelings rather than make claims about specific events.[2] As far as we can determine, the typology encompasses the logically possible situations in which one can test deception within a setting that obeys ethical considerations.[3]

The two highly controlled scenarios are referred to by Miller and Stiff as reaction assessments and Exline procedures.

In reaction assessments, subjects are told to hide their true reactions to scenes they are shown, some of which are pleasant and some unpleasant. Subjects in these experiments are people who are motivated to hide their true feelings in the context of a job situation—for example, medical personnel. The original reaction assessment, designed by Ekman and Friesen (1974), had student nurses view slides of both beautiful landscapes and badly injured burn victims. For each slide the subjects were instructed to say that they were feeling happy, contented, and relaxed, a true claim for the pleasant scenes and a false one for the unpleasant ones. While claims about feelings are usually beyond verification, Miller and Stiff (1993) claim for the unpleasant slides that "we have yet to encounter anyone who was not disturbed by these graphic images" (24), thereby giving the experimenter a basis for ground truth verification.

While reaction assessments provide reliable ground truth and the ability to observe body movements and vocal behavior under true and false conditions, they do not provide the free form extended narrative that is likely to contain a robust number of verbal cues to deception.

The Exline procedure, developed by R. E. Exline and his colleagues (1970), implicates subjects in a cheating or stealing incident, often using a confederate to facilitate the incident. The incident and the subject's behavior during the incident provide the ground truth against which to verify the subject's behavior during the post-incident interview. A skilled interviewer may engage the subject in sufficient narrative to provide a baseline of truthful behavior as well as an adequate number of verbal cues to deception. The motivation to lie, however, is usually problematic for the Exline approach, with motivation including money offered as a reward and/or the threats of discovery and punishment if the cheating or theft is made known. A small monetary reward is an artificial inducement to lie that rarely corresponds to a real-world motivation to deceive, while threats run

into ethical problems, as indeed does the inducement to lie or cheat in artificial circumstances.

The more open-ended scenarios in the Miller and Stiff (1993) typology are the uninterrupted message presentation and interaction analysis. In uninterrupted message presentations, subjects are asked to give a favorable presentation of a position that they endorse and one they do not endorse (Knapp et al., 1974; Newman et al., 2003). Since the subjects' true feelings toward the endorsed and disavowed views are known prior to the presentation, the ground truth is readily available as is the baseline of truthful behavior. These presentations also provide the extended narrative necessary to test verbal cues to deception. However, since subjects are told to lie in the unendorsed presentation, the scenario lacks subject motivation to lie.

Interaction analysis involves the post facto analysis of deception in communicative transactions (Bavelas et al., 1990; Stiff et al., 1992). In interaction analysis, the motivation to lie can be high, depending on the context. Court or government testimony, police interviews and job interviews, for example, all provide venues where deception may be used in the attempt to avoid fines, jail time, or unemployment. All of these venues provide the extended narrative and a period of general questioning that can establish a baseline of truthful behavior.

Interaction analysis involves a real world situation, not a simulated experiment, and satisfies a basic requirement of corpus linguistics for real data. For this reason, we chose real-world interactions—interactions where subjects might be highly motivated to lie—to test the hypothesis that verbal cues correlate positively with deception. However, the establishment of ground truth is a serious issue for this approach. It requires a large amount of detective work, both in searching for internal inconsistencies in the document and in fact checking. In the next two sections, we discuss how the data was gathered and how the ground truth used to test for deception in the data was established.

3. The data

Testing the deception cue hypothesis required two corpora: one to develop the cue tags and the model, and one to test the model and the tags. Both corpora were assembled from publicly available sources, including web sites and police case files.

The corpora used in the study were all narrative intensive, as is required by the interaction analysis approach to deception detection data. The extended narrative on the part of the speaker whose veracity is under question allows for a baseline of truthful behavior against which to compare the deceptive behavior.

The choice of cases was determined entirely by availability. It is critical to obtain original documents for the corpus and, where possible, for sources of ground truth. While many cases are reported on in the press and scholarly works, comparatively few reports come with publicly available documents that are easily

discovered. We relied heavily on Court TV and other web sources, as well as published works and local police documents.

The data includes both speech and written narrative. We received most of the spoken data as transcripts; the interview data, received as recordings, was transcribed on site.

The written statements included in both corpora were produced as parts of a police interview. The purpose of requesting the statement is to obtain an account in the interviewee's own words and to do this before time and questioning affect the interviewee's thinking. Hence the written statement is analogous to a lengthy interview answer, and the language used is much closer to speech than writing, as the following example from an accidental deer hunting death statement shows:

> I shot at the second deer. I couldn't tell if I hit it so I began to follow the tracks. At one point I saw a mark in the snow that could have been blood but I wasn't sure. I sort of became disoriented as I tracked the deer as to where the road was.

The development corpus included narratives from Susan Smith confessing to the murder of her two sons, a deer hunter describing the accidental shooting of a dog walker, a suspect confessing to an armed robbery, Scott Peterson prior to his being charged with his wife's murder, Dr. Jeffrey MacDonald describing the murder of his wife and two daughters before being charged with the crime himself, a murder suspect describing how he spent the weekend while his wife went missing, and several suspects charged with murder in the British Virgin Islands. Information on the development corpus is given in Table 1.

Table 1. Development corpus data.

Case	Case type	Narrative type	Words
Susan Smith	criminal: murder	written statement	602
Deer Hunter	criminal: manslaughter	written statement	603
True_Stickup	criminal: armed robbery	written statement	198
Peterson	criminal: murder	spoken interview	1,3581
MacDonald	criminal: murder	spoken deposition	1,872
False_acct	criminal: murder	spoken deposition	3,674
BVI	criminal: murder	4 spoken depositions	11,029
Total			31,559

This corpus was used to develop a deception detection manual and dictionary that give us explicit lists of words and phrases exemplifying each cue. An excerpt from the dictionary is given in Figure 1.

These lists were used both as reference material for the annotators who were tagging the test corpus and to write software for an automated deception cue tagger.[4]

Word/phrase	Cue type	Comment	Example
a little	hedge	When used as a quantity expr. and as a quantifier on non-count nouns	*{a little%HDG} confused, {a little%HDG} TV (vs. a small TV)*
a little bit/little bit	hedge	similar to *a little*	*He was maybe {a little bit%HDG} taller than me*
a little while	time loss		*I was walking around for {a little while%TL}*
a lot of	hedge		*I sent {a lot of%HDG} information*
a nervous wreck	negative emotion	following *I was*	

Figure 1. Sample from deception cue dictionary.

The corpus used to test the model included three criminal and two civil cases and describes a mix of violent and property crimes, white-collar crime, and civil litigation. While the development corpus included only criminal cases, civil cases were added to the test corpus to determine how generalizable the deception cues would be to the civil cases. The socioeconomic status of the speakers in both corpora ranges from the highly educated to street criminals. Information on the test corpus is given in Table 2.

Table 2. Data used in the experiment.

Case	Case type	Narrative type	Words
Guilty Nurse	criminal: arson	written statement	252
Johnston	civil: sale of tobacco to teens	spoken deposition	12,762
MN Interview 1	criminal: theft	spoken interview	2,282
MN Interview 2	criminal: theft	spoken interview	1,640
Routier	criminal: murder	written statement	1,026
Enron	civil: fraud	spoken congressional testimony	7,476
Kennedy	criminal: manslaughter	written statement	245
Total			25,683

4. Annotating the data for deception

Each document in the experimental corpus was tagged for two factors. (1) linguistic deception indicators marked words and phrases associated with deception, and (2) true/false tags marked propositions that were externally verified.

4.1 Deception indicators

To test our hypothesis, we selected 12 linguistic indicators of deception cited in the psychological and criminal justice literature that could be formally represented and automated in a computational system. The indicators fall into three classes.

1. Lack of commitment to a statement or declaration, where the speaker uses linguistic devices to avoid making a direct statement of fact. Five of the indicators fit into this class: (i) linguistic hedges including non-factive verbs and nominals, e.g., *maybe, I think, to the best of my knowledge*; (ii) qualified assertions, which leave open whether an act was performed, e.g., *I needed to get my inhaler*; (iii) unexplained lapses of time, e.g., *later that day*; (iv) overzealous expressions, e.g., *I swear to God*; and (v) rationalization of an action, e.g., *I was unfamiliar with the road*.
2. Preference for negative expressions in word choice, syntactic structure and semantics. This class comprises three indicators: (i) negative forms, either complete words such as *never* or negative morphemes as in *inconceivable*;[5] (ii) negative emotions, e.g., *I was a nervous wreck*; and (iii) memory loss, e.g., *I forget*.
3. Inconsistencies with respect to verb and noun forms. There are four indicators in this class: (i) verb tense changes, e.g., *I just feel hopeless. I can't do enough. My children wanted me. They needed me. And now I can't help them*;[6] (ii) thematic role changes, e.g., changing the thematic role of a noun phrase from agent in one sentence to patient in another; (iii) noun phrase changes, where different NP forms are used for the same referent, e.g., in the narrative of Dr. McDonald, he describes *my wife* and *my daughter*, but he refers to them as *some people* when he reports their stabbing to the police; and (iv) pronoun changes, which change the referent or omit the pronoun entirely, e.g., Scott Peterson's description of his activities during the time of his wife's murder have no first person reference: *[drove] to the warehouse, dropped off the boat.*

4.2 The hypothesis

The research literature in both psychology and criminal justice looks at correlations between the occurrence of each cue type and the ground truth; for example, Newman et al. (2003) look at the higher occurrence of negative forms in fabricated statements. However, if a narrator is attempting to deceive and the cues leak because of the inability to concentrate on both the message and the mode of presentation—what the literature calls *cognitive overload*—then it is to be expected that deceivers would show a panoply of cue types. Our hypothesis assumes that deceptive speech will show this mix of deception indicators.

Another factor that characterizes the literature on deception is that it more commonly attempts to characterize the narrator as a liar or a truth-teller, rather than to characterize portions of the narrative as true or false. However, liars do

not lie about everything in the narrative.[7] Our hypothesis assumes that a concentration of cues, of whatever type, will correlate with areas of deception in the narrative and the sparse occurrence of cues will correlate with truthfulness. Testing this hypothesis requires narrative data tagged for deception indicators and for ground truth.

4.3 Linguistic annotation

A team of linguists tagged the corpus for the 12 linguistic indicators of deception described above. For each document in the corpus, two annotators assigned the deception tags independently. Differences in tagging were then adjudicated by the two taggers and a third linguist.[8] Tagging decisions were guided by a tagging manual developed by the team during the adjudication sessions. The manual provides extensive descriptions and examples of each indicator. Annotators were not given access to ground truth facts to avoid outside influence on their tagging decisions.

4.4 Determining individual propositions

Independent of the deception indicator tagging, a member of the team identified verifiable claims, or logical propositions, that could be externally verified. While most of the propositions in the corpus are short—they average ten words per proposition—several are considerably longer. The proposition

> What I said—I think what I meant to say is the that data—the raw data that were provided to me by the marketing research department that were collected by what we knew as the Roper data were people only of the age of 18 and over.

is 49 words yet contains a single claim: my data contained only people over the age of 18.

4.5 Establishing ground truth for propositions in the data

External verification of the truth or falsity of each proposition came from supporting materials, including police reports and court documents accepted into evidence and documents judged to be evidence in a congressional hearing. The police reports for the two MN interviews that involved car break-ins contain descriptions of what the cameras mounted in the parking lots recorded as well as the prior records of the suspects.

Statements internal to the narrative were also used as verification if they contradicted other statements made earlier in the narrative. For example, a confession at the end of an interview was used to refute specific claims made earlier in the interview.

The initial verification judgments were made by technical and legal researchers on the project who had not been involved with the deception cue

tagging. The T/F tags were later reviewed by at least one other technical researcher.

The corpus contains 275 verified propositions. Table 3 shows the ratio of word count to verified proposition for each case.

Table 3. Proposition count for each case.

Case	Words	Propositions	Propositions/Words
Guilty Nurse	252	3	.011
Johnston	12,762	83	.006
MN Interview 1	2,282	43	.018
MN Interview 2	1,640	81	.049
Routier	1,026	10	.009
Enron	7,476	45	.006
Kennedy	245	10	.040

The low rate of verified propositions demonstrates the difficulty of establishing ground truth for this real-world data. However, it is interesting to note the relatively high number of verified propositions in the MN police interview data stemming from a good amount of supporting evidence.

Table 4 gives examples of verified propositions in the test corpus.

Table 4. Examples of verified propositions.

Example	True	False
Smokers under 18 were not the market. I did not really study them.		√
These were publicly available data. They were data collected by the University of Michigan.	√	
I'd probably say he was maybe a little bit taller than me.		√
He must have been already been trying to break into the car before me.		√
All right, man, I did it, the damage.	√	
The decision of Mr. McMahon to leave, the decision was totally separate.		√
The particular meeting you're talking about was in Florida, Palm Beach Florida.	√	

5. Testing the ability of the model to predict deception

A core assumption of our approach is that the presence or absence of a cue is in itself insufficient to determine whether the language is deceptive or truthful. Instead, the likelihood of deception depends on the distribution and density of the indicators. Areas of a narrative that contain a clustering of deception indicators

may consist of outright lies or they may be evasive or misleading, while areas lacking in indicator clusters are likely to be truthful.

Given a document tagged with deception indicators, the identification of deceptive and non-deceptive areas is calculated using moving average and word proximity scores. Initial word proximity scores are determined by counting the number of words between the current word and the nearest tag; the tagged material has a score of 0, words adjacent to the tag have a score of 1, and so on. The word proximity score is then recalculated by dividing each score by the number selected for the moving average window. In our case, trial and error motivates a moving average of 22, comprising 10 words preceding the word being scored and 11 words following. Each word in the text receives a derived proximity score. The scores will be low when tags cluster within the range of the moving average window and high when tags are outside the range. Hence, deceptive areas may be defined as areas of text where the word scores fall below a score determined by the statistical model.

Figure 2 gives an example of the proximity scoring. The score of the current word appears between dashes; the score of each word preceding the current word, starting from the nearest neighbor, appears to the left of the current word's score and the score of each word following the current word, again starting from the nearest neighbor, appears to the right of the current word's score. The scores of the entire row are averaged to give the final score for each word.

but	4 3 2 1 0 1 2 3 4 5 6 -- 3 -- 2 1 0 1 0 1 2 3 4 5 --> 2.40909091
it	3 4 3 2 1 0 1 2 3 4 5 -- 2 -- 1 0 1 0 1 2 3 4 5 6 --> 2.40909091
was	2 3 4 3 2 1 0 1 2 3 4 -- 1 -- 0 1 0 1 2 3 4 5 6 6 --> 2.45454545
just	1 2 3 4 3 2 1 0 1 2 3 -- 0 -- 1 0 1 2 3 4 5 6 6 5 --> 2.5
cracked	0 1 2 3 4 3 2 1 0 1 2 -- 1 -- 0 1 2 3 4 5 6 6 5 4 --> 2.54545455
slightly.	1 0 1 2 3 4 3 2 1 0 1 -- 0 -- 1 2 3 4 5 6 6 5 4 3 --> 2.59090909
So	0 1 0 1 2 3 4 3 2 1 0 -- 1 -- 2 3 4 5 6 6 5 4 3 2 --> 2.63636364
ah	1 0 1 0 1 2 3 4 3 2 1 -- 2 -- 3 4 5 6 6 5 4 3 2 1 --> 2.68181818
four	2 1 0 1 0 1 2 3 4 3 2 -- 3 -- 4 5 6 6 5 4 3 2 1 0 --> 2.63636364
to	3 2 1 0 1 0 1 2 3 4 3 -- 4 -- 5 6 6 5 4 3 2 1 0 0 --> 2.54545455
five	4 3 2 1 0 1 0 1 2 3 4 -- 5 -- 6 6 5 4 3 2 1 0 0 1 --> 2.45454545
inches.	5 4 3 2 1 0 1 0 1 2 3 -- 6 -- 6 5 4 3 2 1 0 0 1 2 --> 2.36363636

Figure 2. Example of proximity scoring.

Both *just* and *slightly* are deception indicators (hedges) and therefore receive a score of 0. The word *was*, immediately before *just*, receives a score of 1, the word *it*, two words away from the hedge, receives a score of 2, and so on. Scores to the right are likewise incremented by 1 for each word distance from the deception indicator. Scores for the entire window of 22 words are averaged to give the final score for that word. The score for a proposition is derived by averaging the scores of the words in the proposition.

To test the hypothesis that a concentration of cues will correlate with a proposition being verified as false, we drew from the corpus 275 propositions that could be tested for ground truth; 164 of these propositions, or 59.6%, were verified as false and the remainder as true.

We tested the ability of the model to predict T/F using Classification and Regression Tree (CART) analysis (Breiman et al., 1984) with 25-fold cross-validation and a misclassification cost that penalizes true misclassified as false to counterbalance the greater number of false propositions in the data. Table 5 shows the results of the CART analysis:

Table 5. T/F Classification based on cue concentration.

		Predicted class		
		false	true	% correct
Actual class	false	124	40	75.6
	true	29	82	73.8

The overall accuracy rate of 74.9%—(124+82)/275—demonstrates that our hypothesis, as instantiated in the deception indicator model, identifies deceptive language at a rate significantly better than chance (chi-square $p < 0.005$). Table 6 shows how the model performed on each case in the test corpus.

Table 6. Deception model accuracy on each test corpus case.

Case	**Propositions**	**Correct (%)**	**Ts misclassified**	**Fs misclassified**
Nurse	3	3 (100)	0	0
MN 2	81	66 (80.49)	12	3
MN 1	43	34 (79.07)	9	0
Enron	45	32 (71.11)	13	0
Routier	10	7 (70)	3	0
Kennedy	10	6 (60)	4	0
Johnston	83	49 (59.04)	24	10

Among the large subcorpora, the model performed best on the MN police data despite the fact that there is no clear socioeconomic match for the street criminals involved in these cases in the development corpus, although in terms of narrative genre, they match the Peterson police interview.

The MN data contains a large number of verified propositions, but so does the Johnston data, which scored lowest. However, as Table 3 shows, the number of verifiable propositions as a function of word count is much lower for Johnston than for the MN interviews. The MN narratives contain short, direct answers that stay on topic. Johnston, on the other hand, is a verbose witness and, in the current model, talk not directly related to the facts under question—talk that does not contain deception cues—dilutes the concentration of cues and raises the proposition scores. Another factor in the Johnston narrative is that, unlike the

other narrators, Johnston is not in a high-stakes situation. He is a retired research economist who would not suffer consequences from a problematic testimony.

The model did much better on the Enron subcorpus, the other civil case in the test set. There are many differences between the Johnston and Enron subcorpora to which we might attribute the differences in score—deposition vs. congressional testimony, statistician vs. CEO and the broader range of questions addressed to the CEO, language involving statistical analysis vs. language involving fraud, and a higher rate of deception indicators in Johnston (873/12,762) as opposed to Enron (360/7,476). We are currently exploring these differences with the aim of building different models for different data types.

6. Conclusion

We described the assembling of two corpora of legal narratives that enabled us to test the hypothesis that there is a correlation between specific types of language behavior identified in the forensic literature and deception. The corpora included a mix of spoken dialogue and monologue as well as written statements that were part of a longer interview. A development corpus was used to identify the words and phrases that characterize the specific types of language behavior. A test corpus was used to test the ability of these words and phrases, when concentrated together, to predict a false proposition.

The mixed results among the subcorpora of the test corpus show that accuracy is dependent on the match between the development corpus and the test corpus. This match is difficult to achieve because the cases are chosen in terms of their availability and the amount of extended narrative they can provide while it is often impossible to know in choosing a case how many verifiable propositions it will yield. We are currently seeking an interview situation where ground truth data can be assembled before the interview, enabling us to select only those interviews with substantial background data to annotate for deception.

Improvement of the model will also involve further consideration of the problematic subcorpora, in particular depositions similar to the Johnston data, to determine why the current deception model performs below average on this data.

We also plan to expand the test corpus to include the less well-represented deception indicators. Hedges, negative forms, verb tense changes, pronoun changes, and memory loss are well represented in the test corpus; noun phrase changes, overzealous expressions, and qualified assertions are moderately well represented; but negative emotions, rationalizations of an action, time loss, and thematic role changes are not well represented. The current sparseness of these indicators in the test corpus gives us no way of determining whether they play a role in characterizing deception.

Long-term work will include consideration of other deception indicators from the literature that take the structure of the discourse into account, including topic changes and balance of narrative detail at the beginning, middle, and end of a discourse.

In conclusion, our attempt to use corpora of real-world legal narrative supported by ground truth investigation to test the claim that specific types of language behavior correlate with deception is, as far as we know, unique in the literature on deception. The results of this first attempt, in particular the near 75% accuracy rate in predicting deception, show the feasibility of the approach.

7. Notes

1. Statements about feelings and attitudes are also not useful for the real-world applications we consider.
2. Deception experiments in the United States have to conform to Institutional Review Board standards, which include voluntary, informed consent; protection of privacy and confidentiality; minimization of risk; demonstration that benefit outweighs risk; the protection of vulnerable populations; and inclusion of persons likely to benefit from the research.
3. 75% of the deception indicator tagging is now done automatically.
4. The single word answer *no* is not regarded as a deception indicator since it represents a true negative response.
5. This was part of Susan Smith's statement to the police after she murdered her sons. Only she knew they were no longer alive.
6. We exclude sociopathic liars here.
7. Because the original tagging work was focused on building up the tag bank, inter-rater reliability statistics were not collected. However, current work on new corpora shows a reliability rate of 96%.
8. We used the QUEST program described in Loh and Shih (1997) for the modeling. QUEST is available at <http://www.stat.wisc.edu/~loh/quest.html>.

References

Adams, S. (2002), "Communication under stress: indicators of veracity and deception in written narratives," Ph.D. dissertation, Virginia Polytechnic Institute and State University.

Bachenko, J., E. Fitzpatrick, and M. Schonwetter (2008), "Verification and implementation of language-based deception indicators in civil and criminal narratives," in: *Proceedings of the 22nd international conference on computational linguistics*. ACL, 41–48. Online at: <http://www.aclweb.org/anthology/C/C08/C08-1006.pdf>.

Bavelas, J. B., A. Black, N. Chovil, and J. Mullett (1990), *Equivocal communication*. Newbury Park, CA: Sage Publications.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984), *Classification and regression trees*. London: Chapman & Hall/CRC Press.

DePaulo, B. M., J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper (2003), "Cues to deception," *Psychological bulletin*, 129(1): 74–118.

Ekman, P., and W. V. Friesen (1974), "Detecting deception from the body or face," *Journal of personality and social psychology*, 20: 288–298.

Exline, R. E., J. Thibault, C. B. Hickey, and P. Gumpert (1970), "Visual interaction in relation to Machiavellianism and an unethical act," in: R. Christie and F. L. Geis (eds.) *Studies on Machiavellianism*. New York: Academic Press, 53–75.

Knapp, M. L., R. P. Hart, and H. S. Dennis (1974), "An exploration of deception as a communication construct," *Human communication research*, 1: 15–29.

Miller, G. R., and J. B. Stiff (1993), *Deceptive communication*. Newbury Park, CA: Sage Publications.

Newman, M. L., J. W. Pennebaker, D. S. Berry, and J. M. Richards (2003), "Lying words: predicting deception from linguistic styles," *Personality and social psychology bulletin*, 29: 665–675.

Smith, N. (2001), *Reading between the lines: an evaluation of the scientific content analysis technique (SCAN)*. London, UK: Police Research Series. Online at: <www.homeoffice.gov.uk/rds/prgpdfs/prs135.pdf>.

Stiff, J. B., H. J. Kim, and C. Ramesh (1992), "Truth biases and aroused suspicion in relational deception," *Communication research*, 19: 326–345.

Vrij, A. (2000), *Detecting lies and deceit*. Chichester, UK: Wiley.

Dispersions and adjusted frequencies in corpora: further explorations

*Stefan Th. Gries**

University of California, Santa Barbara

Abstract

In order to adjust observed frequencies of occurrence, previous studies have suggested a variety of measures of dispersion and adjusted frequencies. In a previous study, I reviewed many of these measures and suggested an alternative measure, DP (for 'deviation of proportions'), which I argued to be conceptually simpler and more versatile than many competing measures. However, despite the relevance of dispersion for virtually all corpus-linguistic work, it is still a very much under-researched topic: to the best of my knowledge, there is not a single study investigating how different measures compare to each other when applied to large datasets, nor is there any work that attempts to determine how different measures match up with the kind of psycholinguistic data that dispersions and adjusted frequencies are supposed to represent. This article takes exploratory steps in both of these directions.

1. Introduction

Whether one likes it or not, corpus linguistics is all about distributional data, and virtually every corpus-based paper reports how often a linguistic phenomenon occurred or how often it co-occurred with some other linguistic phenomenon or extralinguistic variable. Such frequency data are used for several different purposes: sometimes they are just used descriptively, but outside of particular traditional schools of corpus linguistics, they are also often used to support particular points or applications in the domains of applied and theoretical linguistics as well as a tool for psycholinguists and psychologists. For example, in some theoretical approaches, such as cognitive linguistics or usage-based grammar, frequency data are now regularly used in the domains of first- and second/foreign-language acquisition, the study of language and culture, grammaticalization, phonological reduction, morphological processing, syntactic alternations, etc.

Interestingly enough, many of these approaches assume a connection between observed frequencies in a corpus and some mental correlate: in first-language acquisition, input frequency is one of the most important determinants of word and construction learning; in cognitive-linguistic approaches, frequency of encounter is one of the central determinants of degree of mental entrenchment/familiarity; for example, observed frequencies (or their logs) are good proxies toward the familiarity of words—see Howes and Solomon (1951) for recognition times, Oldfield and Wingfield (1965) as well as Forster and

Chambers (1973) for naming times, and Ellis (2002a, b) as well as Jurafsky (2003) and Gilquin and Gries (2009) for overviews. Thus, in probabilistic models of language production and comprehension, mental entrenchment in turn is correlated with accessibility (such that, for example, high frequencies of exposure make nodes more available for activation).

Despite this central role of frequency in linguistics in general and in psycholinguistics in particular, it has become clear that for a variety of reasons, frequencies of occurrence are not a perfect predictor of aspects of processing. The first reason is the complexity of all aspects entering into processing effort: no one would deny that the processing of words and concepts is determined by many more though highly intercorrelated aspects such as salience of words and concepts, recency of occurrence, and concreteness/manipulability, to name but a few. Thus, any kind of frequency effect will be ridden with noise and, hence, necessarily indirect. The second reason is that frequency of occurrence, however straightforward to define, does not enter into a straightforward one-to-one relationship with aspects of processing because any particular frequency of occurrence can arise from very different distributional patterns: a word *w* may occur 18–20 times in each of ten very different registers, or it may occur 190 times in only one of the ten registers. While these two results look the same in a frequency list of the complete corpus of ten registers, it is obvious that these results would not be the same: they would not be the same for the corpus linguist who may be interested in register-dependent vocabulary differences, and they would not be the same for the psycholinguist or language acquisition researcher who knows that learning process in general exhibit a distributed learning or spacing effect.

> Given a certain number of exposures to a stimulus, or a certain amount of training, learning is always better when exposures or training trials are distributed over several sessions than when they are massed into one session. This finding is extremely robust in many domains of human cognition. (Ambridge et al., 2006: 175)

Surprisingly, there is not much corpus-linguistic work that deals with or let alone incorporates this potential bias, which in corpus linguistics is referred to as *dispersion*. I know of only a few studies that attempt to address this problem by developing measures of dispersion (i.e., measures that quantify the homogeneity of the distribution of a word in a corpus) or adjusted frequencies (i.e., frequencies that penalize words that are attested only in a small part of a corpus), and there is also only another handful of studies that actually use these measures or study them in more detail. In Gries (2008), I discuss all the measures proposed so far and illustrate that using frequencies alone runs the risk of yielding incorrect results. More specifically, I

- exemplified how frequencies of (co-)occurrence can be quite misleading.

- argued that measures of dispersion as well as adjusted frequencies may be needed to study and characterize our data more accurately.
- suggested a new and intuitively simple dispersion measure called *DP* to address several of the shortcomings that existing measures exhibit.
- provided some resources for researchers: two small computer programs to compute dispersion measures and adjusted frequencies as well as dispersion measures and adjusted frequencies for thousands of words from four different corpora.

(See that study for definitions of and references on all the measures discussed here.) However, it is quite obvious that a variety of issues in this area remains to be explored in more detail, especially given that dispersion characteristics can influence any given corpus-based statistic.

First, we need much more information about the properties of the measures. Lyne's (1985) groundbreaking work is a laudable start: using scatterplots to compare a few dispersion measures, he was the first to try to come to grips with the various ways in which measures differ. However, his study was restricted to the few measures available at that time, and today's computational possibilities allow for much larger and/or more detailed investigations of the measures he used and later ones. For example, little is known about

- how the results of different dispersion measures or adjusted frequencies compare to each other (beyond Lyne's above study). This is problematic since there are now different kinds of measures.
    - First, parts-based measures, which take into consideration how often an element in question is attested in parts of the corpus, but disregard the order of corpus parts as well as where in these parts element in question occurs.
    - Second, distance-based measures, which take into consideration the distances between successive occurrences of the element in question in a corpus (and hence their order), but not its frequencies in different parts of a corpus.
    - Finally, hybrids, which take into consideration both the number of occurrences of the element in question in each corpus part and, within each part, distances between successive occurrences.
- to what degree, if any, these measures come in quantitatively definable, meaningful groups.
- which kinds of distributions (of authentic data) yield what kinds of results.

Such issues are relevant for, for example, being able to better compare dispersion measures from different studies in order to choose the best measure for a task—or at least choose measures that are better suited than other measures for the particular tasks at hand. In Section 2 of this paper, I will therefore compare the behavior of all published measures of dispersions and all adjusted frequencies

I have come across on the basis of the 17,481 most frequent types (10,294,637 tokens) in the spoken component of the British National Corpus World Edition (XML version).

Second, we need to be more serious about validating our dispersion measures and adjusted frequencies. (Strictly speaking, this also applies to measures of co-occurrence strengths, but this is beyond the scope of the present paper.) Devising statistics that are theoretically motivated and make intuitive sense when applied to corpus data is a useful step, although only the first step.

For example, given what we have seen earlier regarding the correlation of observed frequencies (or their logs) and the familiarity of words, psycholinguists or psychologists often use frequency information of words from corpora or databases to create experimental stimuli with the intent to control for frequency or familiarity. However, if dispersion plays the role some corpus linguists have argued, then controlling for frequency alone may turn out to be insufficient unless dispersion is considered at the same time. For corpus linguists, that means that our measures must be validated against corpus-external evidence because, strictly speaking, as long as we corpus linguists do not show that our dispersions and adjusted frequencies correspond to something outside of our corpora, we have failed to provide the most elementary aspect of a new measure—its validation.[1]

How could we provide such evidence? First, we can perform experiments ourselves. For example, one could run experiments (i) on the fictitious corpus distributions discussed in Lyne (1985) and Gries (2008) to determine whether the measures are able to distinguish them or not and (ii) to determine which measures' results from large balanced corpora are most compatible with subjects' intuitions regarding the words' overall centrality in a language. Since dispersion and adjusted frequencies are used as proxies to familiarity, one could also check whether ways of presenting children with nonce words that differ in terms of the dispersion patterns of the word in question lead to different degrees of learning success (see studies on distributed learning such as Ambridge et al., 2006). Thankfully, the number of experimental validations of corpus-based studies is steadily increasing, and the field of dispersion should be no exception to this general trend.

Second, we can reanalyze published psycholinguistic data. In Section 3 below, I will correlate dispersion measures and adjusted frequencies with the response time latencies of Spieler and Balota (1997) and Balota and Spieler (1998), as well as lexical decision task data from Baayen (2008).

2. Dispersions and adjusted frequencies: intercorrelations

To explore how dispersion measures and adjusted frequencies are intercorrelated with each other, I used data from the spoken component of the British National Corpus (BNC) World Edition (XML). More specifically, I wrote an R script that

- loaded each corpus file of the BNC World Edition (XML) that contained spoken data, converted it to lower case, and retained only the lines that contained sentence numbers (regular expression: "<s·n").
- deleted all sequences of nonword tags and the material they refer to (regular expression: "·*<(?!w·.*?>).*?>[^<]*").
- split up the remaining data at sequences of optional spaces and word tags (regular expression: "·*<w.*?>").
- printed the resulting word list into an output file such that the file contained all the word tokens from the corpus in the order of the files followed by one tab stop followed by the name of the file in which the word occurred.[2]

Then, for all word types that occurred ten or more times in the corpus, I used another R script to compute all 29 dispersion measures and all adjusted frequencies discussed in Gries (2008); the corpus parts I assumed were the individual files, which were processed in alphabetical order of their filenames. As a result, I obtained a table with these dispersion measures and adjusted frequencies in the columns for the 17,481 word forms in the rows.[3, 4]

Intercorrelations between these measures were explored using hierarchical agglomerative cluster analyses and, for additional graphical exploration, principal component analysis. Hierarchical agglomerative cluster analysis is a statistical tool well-suited to the task at hand. Such cluster analyses try to find structures in the data by successively amalgamating individual measures into larger groups such that the within-groups similarities are as large as possible compared to the between-groups similarities. While such a clustering approach is bottom up and data driven, the researcher has to make at least two important decisions. First, one has to decide on a measure of pairwise similarity on the basis of which the different elements—the dispersion measures and adjusted frequencies—are compared to each other. Second, one has to decide on an amalgamation rule, an algorithm that determines how groups of elements are merged. As to the former, I use a fairly standard measure, namely 1-*r*, where *r* is the Pearson product-moment correlation between the vectors of any two measures that are being compared.[5] As to the latter, I use Ward's method because it has been shown to be a reliable measure and good at identifying small clusters, and because of its affinity to the logic underlying ANOVAs. To avoid distortions by the different scales of the measures, the values were *z*-standardized by column,[6] and I did separate analyses for the dispersion measures and the adjusted frequencies.

2.1 Dispersion measures

The result of the cluster analysis on dispersion measures is shown in Figure 1 (the abbreviations of the measures are listed in Appendix 1).

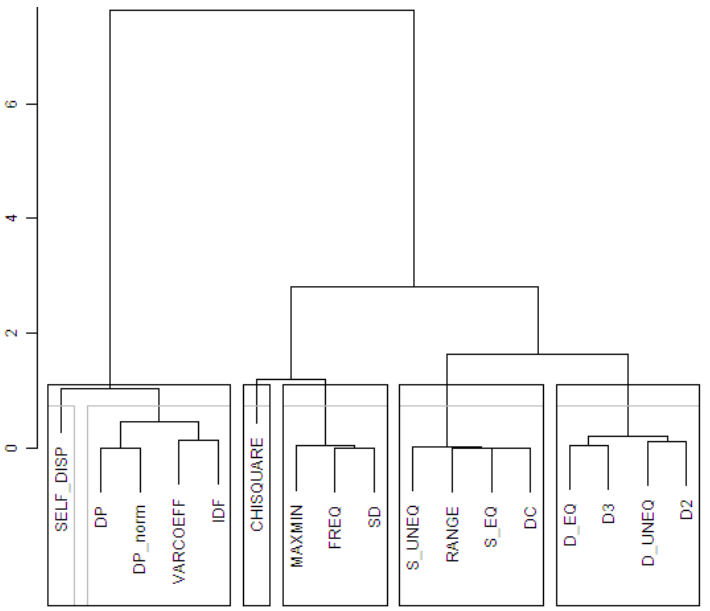


Figure 1. Dendrogram of dispersion measures.[7]

The results suggest five different clusters:

- The maximal average silhouette width for a cluster solution (0.73) was obtained for six clusters (see the grey boxes), but this comes at the cost of assuming two clusters that consist of only one measure (assuming one-measure clusters is undesirable because such clusters mean that the one measure is in fact unique and cannot be merged with another one, which is after all the whole point of clustering).
- The second highest average silhouette width is practically the same (0.729) but has only one single-element cluster and a much higher average silhouette width than the next solution with fewer clusters (four clusters: 0.675).
- The principal component analysis returned only four principal components with eigenvalues larger than 1; the loadings of the first two principal components (the first on the *x*-axis, the second on the *y*-axis), which together account for 77.2% of the variance in the data, are plotted in Figure 2 (the polygons represent the clusters from the HCA).

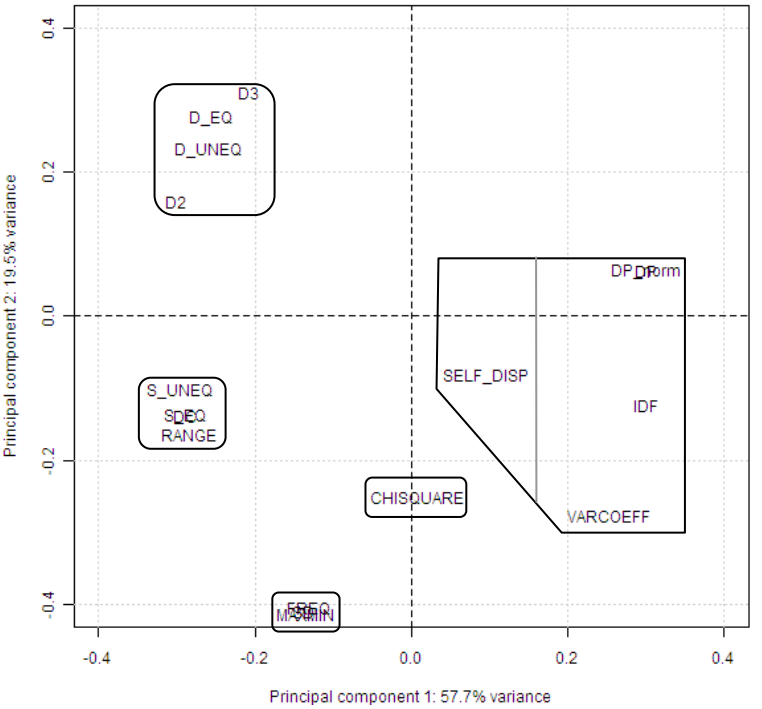


Figure 2. Loadings of the first two principal components.

For reasons of space, I cannot discuss the results here in great detail. It is obvious, however, that the proposed 16 dispersion measures constitute five or six different kinds of measures that differ mainly along two dimensions and that exhibit varying degrees of homogeneity: both versions of Rosengren's *S*, *DC*, and the *range* are very similar to each other whereas the cluster containing *idf* and the variation coefficient is rather heterogeneous. Interestingly, a measure such as Carroll's D_2, whose creator harshly attacked Juilland et al.'s *D* for a variety of perceived shortcomings, is actually very similar to it in terms of its overall behavior—in fact, much more similar than to most other measures. In addition, I checked how each cluster relates to raw frequency by inspecting one central member of each cluster. In order to force all values into a comparable range and give them the same orientation (high values indicate high clumpyness and low values indicate more even distributions), I *z*-standardized

- the vectors of 1-*DC* value and 1-D_{equal}.
- the vectors of *idf*, chi-square and *DP* values.

These values were then plotted against log frequency and summarized with smoothers.[8] The result shows that the different groups of values behave very differently with respect to frequency; see link 2 in Appendix 2. *DC*, D_{equal}, *idf*,

and *DP* become smaller with larger observed word frequencies, but, on the whole, chi-square does not. In addition, the measures exhibit quite different ranges: while *DC* and *DP* are fairly similar, the D_{equal} has more larger values than *idf*, and chi-square has many large values. Finally, the curvature of the smoothers sometimes differs considerably: *DC* and *DP* again behave similarly but different from both D_{equal} and *idf*.

What does all this show? In the present form, the data do not show much in terms of specific content. What they do show, however, is that different measures of dispersion will yield *very* different (ranges of) values when applied to actual data. Researchers must exercise caution in their choice of a measure of dispersion for their data: not only should they make sure that they choose a measure that exhibits all of the theoretically desirable characteristics,[9] but they might also want to consider reporting or basing their subsequent analysis on the results of more than one measure, ideally from measures from the different clusters represented in Figures 1 and 2. Interestingly, the one dispersion measure that is conceptually very different from all others does not exhibit a particularly special status in the evaluation: Washtell's self-dispersion is the only measure that not only takes into consideration the number of times an element is observed in a corpus part, but also the distances between the occurrences. On the one hand, this may seem like a theoretically very attractive feature, but it can also be applied only when a word occurs more than once in a corpus part. However, while this measure is different enough to constitute a cluster on its own when the less parsimonious six-cluster solution is adopted, the principal component analysis shows that it is located in a relatively populated area of principal component space. More and maybe more diverse data are required to shed light on whether or not the additional computational effort of this theoretically attractive feature of self-dispersion is justified.

2.2 Adjusted frequencies

Interestingly, the result of the cluster analysis on adjusted frequencies does not merit a figure. Apart from Kromer's U_R, all other measures are grouped together; the only one that may be a little bit different from the rest is f_{AWT}: the average silhouette width for the two-cluster and three-cluster solutions are 0.89 and 0.65 respectively. What is more interesting to note is that the two different kinds of adjusted frequencies are not distinguished very much: the distance-based measures proposed by Savický and Hlaváčová are grouped together with several parts-based measures, which disregard distance information. Also, when one looks at how much in percent each adjusted frequency reduces the actually observed frequency, then the three measures that are farthest away from each other in the dendrogram behave completely differently; see Figure 3, which shows the non-parametric smoothers for Rosengren's *AF* (for equal corpus parts), Kromer's U_R, and Savický and Hlaváčová's f_{AWT}, and link 3 for the complete plot.

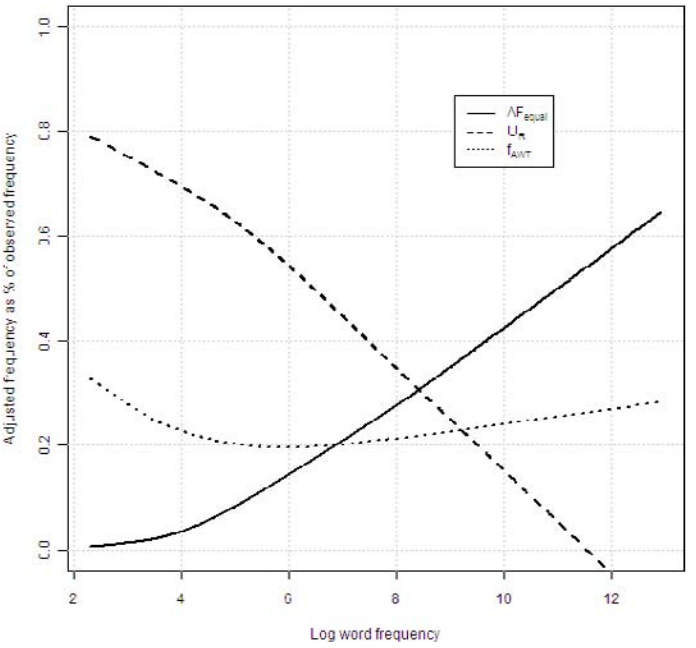


Figure 3. Nonparametric smoothers summarizing three adjusted frequencies.

It is hard to imagine a more diverse result. The more frequent words are, the less Kromer's U_R reduces their frequency, but at the same time the more Rosengren's *AF* does, and f_{AWT} is different from both. I have little to say about this particular result other than that it clearly emphasizes that we know next to nothing about how different adjusted frequencies behave and what they actually mean or do. More exploration is necessary but even more important is that we begin to validate the two dozen or so dispersion measures and adjusted frequencies we have at our disposal. A first step in this direction will be taken in the next section.

3. Dispersions and adjusted frequencies: validation against psycholinguistic data

While dispersion measures and adjusted frequencies were developed with rather practical motivations in mind (e.g., to provide lexicographers with more reliable statistics than raw frequencies), it is probably fair to say that our knowledge of dispersion measures and adjusted frequencies is approximately inversely proportional to what we know about their accuracy, reliability, and predictive power. In this section, I want to briefly explore how the measures we have relate to the kind of psycholinguistic data they are presumably supposed to relate to. If dispersion measures are really better indicators of, for example, the familiarity of words (and, hence, somewhat indirectly to maybe even to the concepts these

words evoke), if adjusted frequencies are truly more appropriate indicators of cognitive entrenchment, then we should find robust correlations between our measures and psycholinguistic results such as response time latencies. Unfortunately, it will become obvious that the data raise more questions than they answer.

As a first example, I correlated the response time latencies of young and old native speakers of English to monosyllabic words from Spieler and Balota (1997) and Balota and Spieler (1998). All dispersion measures and adjusted frequencies were centered and the correlation coefficient used was Kendall's τ. Given that not all dispersion measures have the same orientation (see Section 2.1 above), the correlations between the measures and the response time latencies can be both positive and negative: the larger an effect, the more Kendall's τ will deviate from zero; see Figure 4 for the results for young speakers and Figure 5 for the results for old speakers; the *x*-axis labels (*d* and *f* indicate whether the plotted measure is a measure of dispersion or an (adjusted) frequency and given the large *n*, all correlations are highly significant.

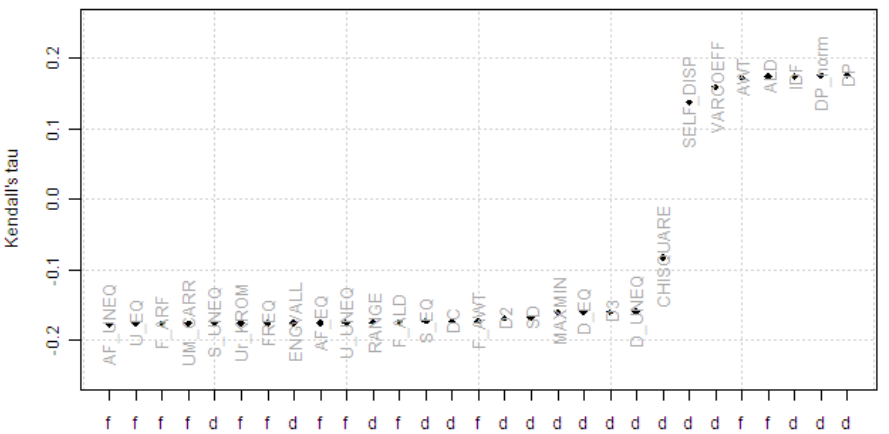


Figure 4. Correlations between Balota and Spieler's (1998) response time latencies (young speakers) and the dispersion measures and adjusted frequencies surveyed in Gries (2008).

In some sense, the results are striking. On the one hand, both panels show the same measures as resulting in the strongest correlations: *ALD*, *DP*/*DPnorm*, and *idf* (as measures with positive correlation coefficients) and AF_{uneq} and U_{uneq} (as measures with negative correlation coefficients). On the other hand, it is equally obvious that with very few exceptions, it doesn't seem to matter which measure is chosen since most of the correlations are of the same strength (which also means that Kromer's U_R—despite the claim of it being more psycholinguistically grounded—does not result in a stronger correlation with the psycholinguistic data).

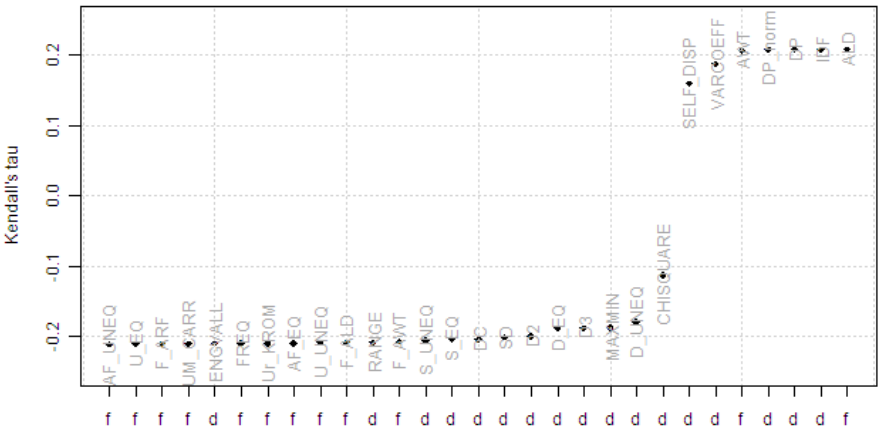


Figure 5. Correlations between Balota and Spieler's (1998) response time latencies (old speakers) and the dispersion measures and adjusted frequencies surveyed in Gries (2008).

Even this interim conclusion, however, is undermined once we do the same kind of computation for the lexical decision task data of Baayen (2008), which are represented in Figure 6.

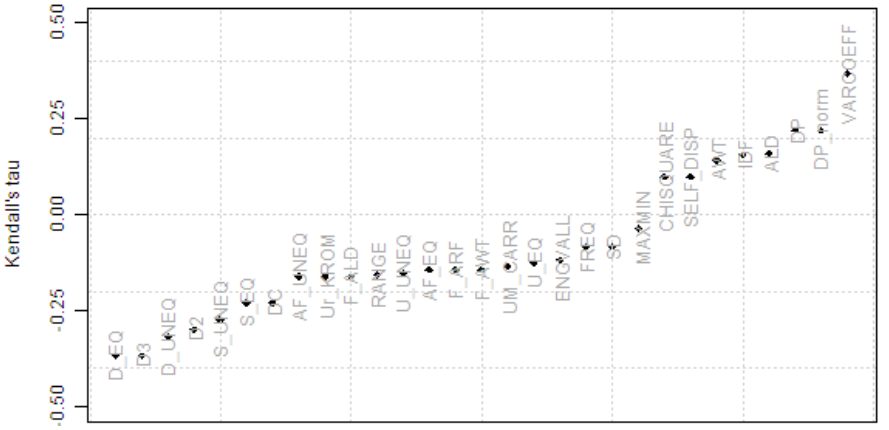


Figure 6. Correlations between Baayen's (2008) lexical decision task times (for native speakers) and the dispersion measures and adjusted frequencies surveyed in Gries (2008).

Again, *ALD* and *DP*/*DP*$_{norm}$ are among the strongest correlations, only surpassed, perhaps surprisingly, by the variation coefficient, but D_{equal} and D_3 also exhibit strong correlations although their correlations with Balota and Spieler's data were only somewhat moderate. On the more positive side, compared to Figures 4 and 5, this time there is a distinct cline with some measures clearly very

close to a null correlation, and as a matter of fact, only D_{equal}, D_3, $D_{unequal}$, and the variation coefficient correlate significantly with the psycholinguistic measure.

4. Summary and some (preliminary) conclusions

While the results of Section 2 provide at least some clue(s) for future studies, the results of Section 3 do not yet inspire a lot of hope. Section 2 showed that when the proposed dispersion measures are applied to most of the words in the spoken component of the BNC, they fall into approximately five groups along two dimensions and take on a bewildering range of values. It is probably safe to say that chi-square is not a particularly useful measure since across the full range of observed frequencies, it exhibits an extremely high range of values, so chi-square does not appear to be particularly discriminatory. However, apart from that, the dispersion measures differ mainly in the degree to which they reach higher values with increasing frequency, and none of them reaches really high levels of predictive power, which was to be expected (recall Section 1).

For the adjusted frequencies, the picture is more diverse: the measures do not fall into nicely distinct groups other than U_R vs. the rest, but if three core measures are explored, the ways in which an adjusted frequency reduces the observed frequency exhibit all possible directions of correlation with the actually observed frequency.

Despite the diversity of these results, recommendations for future work are clear: avoid chi-square, use several different measures of dispersion from the identified groups, bear in mind the potentially confounding factor of corpus part sizes, and explore self-dispersion as well as the distance-based measures to determine whether or not they ultimately yield more revealing results.

Section 3 brings good news and bad news. The good news is that for all three psycholinguistic measures, there are significant correlations between at least some dispersion measures and adjusted frequencies, the highest absolute correlations are provided by a small set of measures (*ALD, DP,* and the variation coefficient are among them), and, crucially, these measures are more highly correlated with the psycholinguistic results than raw frequencies of occurrence. It is particularly interesting that a general measure of dispersion such as the variation coefficient, which has not been designed specifically to handle corpus data, scores so well. The bad news, however, is that the data are as yet too small and too heterogeneous to allow making more meaningful recommendations than that, (i) focusing on these three measures probably increases the likelihood of good results and (ii) we need to know more.

On a methodological level, it also emerges that even though Lyne's earlier work on comparing different measures of dispersion has been a major milestone, it is now time to include more measures and adopt a multivariate perspective. Lyne used a plot-based exploration on selected (fictitious) distributions, but the present approach shows that (i) looking at more than 17,000 word types and (ii) using more sophisticated methods—robust smoothing approaches, hierarchical

cluster analysis, and principal components analysis—have more to offer than was available at the time of Lyne's work. While this paper could only take a small step toward answering all the questions that arise from the literature, I hope I could provide some initial and interesting results and some incentive to explore these issues further. After all, what are dispersions and adjusted frequencies good for when we don't know what they do and what exactly they measure?

5. Notes

* I thank three anonymous reviewers for their comments and suggestions. The usual disclaimers apply.

1. One laudable exception is the recent work by Ellis (2002a, 2002b), which shows that range has significant predictive power above and beyond raw frequency of occurrence, and it is this kind of evidence we must provide in order to show our efforts are more than devising clever equations.
2. All retrieval operations, statistics, and graphs were computed with R 2.8.0 (see *R* Development Core Team, 2008).
3. Scripts to compute dispersion measures and adjusted frequencies as well as dispersion measures and adjusted frequencies for words from four different corpora are available from my web site; see link 1 in Appendix 2.
4. Since it is as yet an unresolved question exactly how dispersions and adjusted frequencies react to different numbers of corpus parts (especially in combination with differently sized corpus parts), it needs to be mentioned how similar the corpus parts are to each other. In this case, the file sizes (in words) were all rather similar to each other: the relative entropy of the file sizes is 0.914 and thus relatively close to the theoretical maximum.
5. I used 1-r as a measure to be able to better compare the results of the cluster analysis with the of the principal components analysis. A cluster analysis based in Kendall's τ as a similarity measure yielded a virtually identical dendrogram, the sole difference being that the two clusters on the right of Figure 1 were more similar to each other.
6. To *z*-standardize a value *x* from a vector/range of values, you subtract the mean of all the values from *x* and divide the result by the standard deviation of all the values:

$$z = \frac{x - \mu}{\sigma}.$$

7. See the appendix for the meanings of the abbreviations.
8. I used locally weighted polynomial regressions as smoothers, (i.e., regression lines that try to summarize the cloud of points in a scatterplot without the restriction of typical linear regressions that the line must be straight); see ?lowess in R.

9. These "theoretically desirable characteristics" include the ability of dispersion measures (i) to handle differently-sizes corpus parts, (ii) to fall only into the range the dispersion measure is supposed to fall into, (iii) to exhaust that range (i.e., not cluster only in small range of the complete theoretical range), (iv) to not be overly sensitive to the overall numbers of corpus parts, (v) to be sensitive enough, but not too sensitive, given extreme distributions and zero occurrences, and others; see Gries (2008: Section 2.4 and 5 for discussion and exemplification).

References

Ambridge, B., A. L. Theakston, E. V. M. Lieven, and M. Tomasello (2006), "The distributed learning effect for children's acquisition of an abstract syntactic construction," *Cognitive development*, 21: 174–193.

Baayen, R. H. (2008), *Analyzing linguistic data: a practical introduction to statistics with* R. Cambridge: Cambridge University Press.

Balota, D. A., and D. H. Spieler (1998), "The utility of item level analyses in model evaluation: a reply to Seidenberg and Plaut," *Psychological science*, 9(3): 238–240.

Ellis, N. C. (2002a), "Frequency effects in language processing and acquisition: a review with implications for theories of implicit and explicit language acquisition," *Studies in second language acquisition*, 24: 143–188.

Ellis, N. C. (2002b), "Reflections on frequency effects in language acquisition: a response to commentaries," *Studies in second language acquisition*, 24: 297–339.

Forster, K. I., and S. M. Chambers (1973), "Lexical access and naming time," *Journal of verbal learning and verbal behavior*, 12: 627–35.

Gilquin, G., and St. Th. Gries (2009), "Corpora and experimental methods: a state-of-the-art review," *Corpus linguistics and linguistic theory*, 5(1): 1–26.

Gries, St. Th. (2008), "Dispersions and adjusted frequencies in corpora," *International journal of corpus linguistics*, 13: 403–37.

Howes, D. H., and R. L. Solomon (1951), "Visual duration threshold as a function of word-probability," *Journal of experimental psychology*, 41: 401–410.

Jurafsky, D. (2003), "Probabilistic modeling in psycholinguistics," in: R. Bod, J. Hay, and S. Jannedy (eds.) *Probabilistic linguistics*. Cambridge, MA: MIT Press, 39–96.

Lyne, A. A. (1985), "Dispersion," in: *The vocabulary of French business correspondence*. Geneva, Paris: Slatkine-Champion, 101–124.

Oldfield, R. and A. Wingfield (1965), "Response latencies in naming objects," *Quarterly journal of experimental psychology*, A 17: 273–281.

R Development Core Team (2008), R*: A language and environment for statistical computing*. Vienna, Austria: *R* Foundation for Statistical Computing.

Spieler, D. H., and D. A. Balota (1997), "Bringing computational models of word naming down to the item level," *Psychological science*, 8: 411–416.

Appendix

Appendix 1.

Abbreviation	**Measure**
FREQ	observed frequency of word *w*
RANGE	number of parts with word *w*
MAXMIN	max. freq. of *w*/part—min. freq. of *w*/part
SD	standard deviation of frequencies
VARCOEFF	variation coefficient of frequencies
CHISQUARE	chi-square value of the frequency distribution
D_EQ	Juilland et al.'s *D* (assuming equal parts)
D_UNEQ	Juilland et al.'s *D* (not assuming equal parts)
D2	Carroll's D_2
S_EQ	Rosengren's *S* (assuming equal parts)
S_UNEQ	Rosengren's *S* (not assuming equal parts)
D3	Lyne's D_3
DC	Distributional Consistency
IDF	Inverse Document Frequency
ENGVALL	Engvall's measure
U_EQ	Juilland et al.'s usage coefficient *U* (assuming equal parts)
U_UNEQ	Juilland et al.'s usage coefficient *U* (not assuming equal parts)
UM_CARR	Carroll's U_m
AF_EQ	Rosengren's Adjusted Frequency *AF* (assuming equal parts)
AF_UNEQ	Rosengren's Adjusted Frequency *AF* (not assuming equal parts)
Ur_KROM	Kromer's U_R
F_ARF	Savický and Hlaváčová's f_{ARF}
AWT	Savický and Hlaváčová's *AWT*
F_AWT	Savický and Hlaváčová's f_{AWT}
ALD	Savický and Hlaváčová's *ALD*
F_ALD	Savický and Hlaváčová's f_{ALD}
SELF_DISP	Washtell's self-dispersion
DP	Gries's Deviation of Proportions
DP_norm	Gries's Deviation of Proportions (normalized)

Appendix 2.

Link 1:
<http://www.linguistics.ucsb.edu/faculty/stgries/research/dispersion/links.html>

Link 2:
<http://www.linguistics.ucsb.edu/faculty/stgries/research/dispersion/comparison_dispersion.png>

Link 3:
<http://www.linguistics.ucsb.edu/faculty/stgries/research/dispersion/comparison_adjfreq.png>

Probabilistic tagging of minority language data: a case study using Qtag

Christopher Cox

University of Alberta

Abstract

While probabilistic methods of part-of-speech tag assignment have long received consideration in corpus and computational-linguistic research, less attention would appear to have been paid to date to the development of tagging accuracy over rounds of iterative, interactive training in applications of these methods. Understanding this aspect of probabilistic tagging is arguably of particular importance to the successful construction of minority language corpora, where financial resources for corpus development are often limited and no fixed standards for either orthography or part of speech assignment may necessarily exist. This paper therefore presents a case study in the application of pure probabilistic tagging, as represented by Qtag (Tufis and Mason, 1998), to minority-language data from Mennonite Low German (Plautdietsch). Concentrating upon the relationship of several factors (including training data size, tag set complexity, and orthographic normalization) to the development of tagging accuracy, the present study conducts computational simulations of the iterative, interactive training process to compare the interactions of these factors quantitatively over time. The study concludes with a discussion of these factors' relevance to the development of accuracy in tagging as well as of potential confounds to the application of probabilistic tagging methods to similar minority language data.

1. Introduction

Probabilistic methods in the assignment of grammatical category labels to natural language data have long represented an area of active research in computational linguistics (see Church 1988, DeRose, 1988). Such methods, while certainly not without deterministic counterparts, have arguably been of particular importance in the recent history of corpus linguistics, where their application in large-scale corpus construction projects has often met with considerable success. This has provided researchers within both computational and corpus linguistics with quantities of tagged natural language data beyond historical parallel, and thus, in part, fostered the development of quantitative methods of linguistic modelling which make active recourse to computationally assigned attributes of their primary data.

The application of probabilistic part-of-speech assignment methods has proven profitable in many such large-scale corpus construction projects, where time, linguistic data, technical expertise, and financial resources are often comparatively abundant. Perhaps less documented, however, are the challenges of applying similar techniques when one or more of these resources is limited. Such

is commonly the case in minority-language corpus construction (see McEnery and Ostler, 2000), where both linguistic and technical-financial hurdles must often be overcome. Standards for the representation of language materials, whether in the form of codified orthographies or of descriptions of grammatical categories which must be present in any systematic tagging of the language at hand, may not have been proposed, sufficiently elaborated, or uniformly adopted within the speech communities represented, thus posing potential problems for the straightforward incorporation of available linguistic resources into an internally consistent corpus. Likewise, from a technical perspective, existing computational techniques proposed for part-of-speech tagging may falter on or simply not be amenable to the typological structure of certain languages. Depending upon the degree of detail sought in tagging, a language with polysynthetic morphology, for instance, may prove a more difficult target for dictionary-based methods of tagging, which rely to varying extents upon the recurrence of word-tag pairs to achieve accuracy, than the primarily morphologically isolating or analytical languages which have featured most prominently among large-scale corpus construction projects to date. In all such cases, limited resources for corpus construction may concomitantly limit the development of language-appropriate computational techniques or linguistic standards which might be applied to the available corpus material.

Where financial and technical resources are limited, then, it would seem important to understand which factors bearing upon corpus development might be expected to produce acceptable results and minimize overall expenditure of effort, given a set of assumed goals and available resources. The present paper offers one such evaluation of the efficacy of probabilistic computational techniques in part-of-speech tagging when applied to minority language data. This assumes the form of a case study in the use of a freely available probabilistic part-of-speech tagger, Qtag (Tufis and Mason, 1998), in the annotation of a small (approximately 120,000 token) corpus of written Mennonite Low German (*Plautdietsch*). Among its other benefits, the adopted case study format presents an opportunity to consider in concrete detail several problems commonly faced in tagging minority-language data and, thus, provides a chance not only to assess the tagging procedures adopted in this particular instance but also, through post-hoc simulation of different tagging models, to discuss alternatives which may have produced comparable results with reduced expenditure of resources. In this way, the present study seeks to offer both a description and assessment of minority-language corpus development "in action" as well as methods of corpus development evaluation which might be of broader use in similar minority-language corpus development projects.

2. Constructing a corpus of *Plautdietsch*

Plautdietsch (ISO 639-3: pdt), the language of the corpus described here, is an Indo-European language of the Germanic subgroup, formerly spoken in the area

of the Vistula Delta in northern Poland. Recent estimates place the number of Plautdietsch speakers globally between 300,000 and 500,000 (see Epp, 1993: 102–3; Epp, 2002; Gordon, 2005), although substantial variation is noted in the relative vitality of the individual speech communities which comprise this estimate. Most numerous among the present-day speakers of Plautdietsch are the descendents of Dutch-Russian Mennonites, an Anabaptist Christian denomination that originated in the Protestant Reformation. As a result of the persecution, emigration, and exile of Dutch-Russian Mennonites over the course of four centuries, sizeable Plautdietsch speech communities are to be found today on four continents and in no fewer than a dozen countries. Considerable dialectal variation is observed between disparate groups of Plautdietsch speakers distinguished historically by differing patterns of emigration and divided contemporarily by both geographical distance and several centuries of mutual isolation. For further discussion of the linguistic history and characteristics of this rather exceptional member of the Germanic language group, refer to the existing literature concerning Plautdietsch. In particular, Epp (1993) presents a thorough overview of the development of Plautdietsch in historical-linguistic context, while the origins of those written varieties represented in the present corpus are discussed in Loewen and Reimer (1985).

Much as in the construction of any other corpus, the development of a corpus of written Plautdietsch must arguably consider not only features of the available linguistic material (e.g., its representativeness within the writing traditions maintained by the contributing speech communities, the varietal features it exemplifies, etc.) and aspects of its digital representation (e.g., the selection of appropriate standards for text encoding and the representation of document structure and linguistic annotations), but also the anticipated users of the final corpus and the uses to which it might be put. The necessity of such planning is further underscored, as the previous section has suggested, when resources for corpus development are comparatively limited. This planning, however, poses problems for the would-be corpus developer: precisely which features should receive immediate attention in corpus construction, and which should be set aside as areas for future development? Which subset of the potential uses and users of the finished corpus should be selected as the specific focus of short-term development? Indeed, even limiting such considerations to the task of probabilistic part-of-speech annotation, selecting which linguistic features to annotate and to what level of detail may be a less-than-trivial undertaking and have serious consequences for the ultimate success or failure of a corpus development project to meet its stated goals, even when other linguistic and technical aspects of such a project are relatively well understood.

In this case, the present corpus of written Plautdietsch, while expected to be suitable for many possible linguistic analyses and acceptable to the speech communities whose language it represents, is intended primarily for research into the syntax of verbal complementation. For the purposes of such investigation, then, it is desirable to have detailed tagging which captures each inflectional category of finite verbs—their tense, person, number, and so on. Given the degree

of dialectal variation found between individual varieties of Plautdietsch, it may also be reasonable to suspect that the varieties themselves represent potentially relevant predictors of variation in verbal syntax. Thus, dialectal attributes should also likely be represented in some form in the corpus. From a technical perspective, the inclusion of these features in the final corpus is made feasible with computational resources for corpus construction furnished largely by the Text Analysis Portal for Research (TAPoR) laboratory at the University of Alberta. The linguistic goals of this corpus development project—that is, to produce a corpus appropriate for quantitative investigations of verbal syntax—are thus relatively clear, and institutional support significantly lessens the technical and financial burdens that might otherwise discourage such an endeavour. Nevertheless, the time required to develop such a corpus is shared with that of later research into verbal syntax, and the management of this resource is important to the success of both projects. Time investment in corpus construction should therefore be minimized wherever possible.

While the benefits provided by institutional support are considerable and should not be underestimated, the development of a corpus of Plautdietsch nevertheless faces several challenges commonly encountered in minority-language corpus construction. The present discussion limits its attention to three such challenges in particular:

1. *No single orthographic standard exists for Plautdietsch*. Both the relatively short history of Plautdietsch as a written language (see Epp 1996: 3) and the geographical dispersion of Plautdietsch speakers globally have contributed to the wide range of orthographic systems and conventions attested in contemporary texts. Spelling systems may vary not only between individual authors, but even between the individual works of a single author, with several noted Plautdietsch writers having elaborated upon their own orthographic standards over the period of several decades (see Nieuweboer, 1999). Thus, Plautdietsch spelling systems may vary where the represented varieties themselves presumably do not (e.g., in the pronunciations of a single author who has developed a spelling system over time) or, in the case of phonetically close orthographies, present potentially valuable sources of information on phonological variation and speakers' perceptions thereof.

   While common conventions for the representation of certain phonemes have emerged across many such spelling systems, this fact in itself does not present an immediate remedy to many of the problems encountered with the presence of multiple orthographic standards in a single corpus. Diversity in spelling is unlikely to cause probabilistic tagging procedures to fail entirely; it is likely, however, to increase the difficulty of inductively training an effective probabilistic model with lexically conditioned probabilistic tagging systems, as the number of orthographically distinct word forms grows with each new spelling system represented in the corpus, and thus increases the overall number of types

which the tagging system must either come to recognize or learn to predict effectively. Perhaps more problematically, having multiple spelling systems represented as such in the corpus renders exhaustive search and retrieval difficult, if not impossible. For a task as basic as retrieving all instances of a given word (to compute a lexical frequency or dispersion measure, for instance), one must essentially search for each possible spelling of that word, a task that would require a priori knowledge of each spelling system in use in the corpus (while making the somewhat generous assumption that each system has been applied consistently and without significant variation in each source work). The magnitude of this problem becomes all the more apparent when attempting to search for pairs or sequences of collocates in the corpus, with a combinatorial increase in the potential number of orthographic variants which must be taken into account in each search.

In addition to the technical challenges posed by variation in spelling, the choice of orthographies remains an issue of some contention among authors of Plautdietsch. Individual writers and publications often express strong preferences for particular orthographic standards or conventions. If the final corpus is to be considered acceptable by the larger Plautdietsch speech community, the orthographies chosen by those authors whose works are represented in the corpus must likely be preserved in some form in the corpus.

2. *No corpora of Plautdietsch have been published to date*. While several studies of Plautdietsch (e.g., Klassen, 1969; Hooge, 1973) have made reference to private corpora assembled largely from independent fieldwork, excerpts of which have occasionally been published in edited form (see Klassen, 1993), no publicly available digital corpora of Plautdietsch exist. No systems of conventions for representing part-of-speech categories ('tagsets') have been proposed for this language, nor indeed is there significant consensus among existing grammatical descriptions of Plautdietsch varieties as to the grammatical categories which must be present in any adequate representation of the language. (Even relatively basic features of the language remain heavily disputed, such as the number of distinct cases in Plautdietsch for which determiners and adjectives inflect, rendering their codification in a standardized tagset more difficult.)
3. *Dialectal variation*. As was noted earlier, substantial variation exists both between and within national varieties of Plautdietsch in their lexical and morphosyntactic features. Whereas the former category of lexical variation is unlikely to prove problematic for probabilistic tagging—word forms characteristic of one or another particular variety will appear as types of limited dispersion in the corpus—the latter category of morphosyntactic variation poses potential difficulties for the training of probabilistic taggers and effective search and retrieval within the finished corpus. The reasons for this are much the same as those for problems related to

orthographic variation: morphological variation in the realizations of common inflectional features (e.g., the form of the regular nominal plural suffix *–e[n]* or of the infinitival verb suffix *–e[n]*) causes an increase in the overall type count in the corpus, presenting a greater number of unique word forms with which the probabilistic tagger must grapple and for which the corpus user must know to search. These problems are only compounded in an orthographically unnormalized corpus: searching for all occurrences of a given word in a dialectally diverse and orthographically unnormalized corpus must take into account not only all possible dialectal variants of that word, but also all possible spellings of each such variant. While perhaps feasible for certain simple lexemes, this solution quickly becomes intractable as the length of the search and the orthographic and dialectal variability of the sought-after tokens increase.

Given these challenges, then, a three-stage construction procedure was adopted for the present corpus of Plautdietsch which was intended to address each of the above challenges in turn. These stages are as follows:

1. *Orthographic normalization*. A separate version of each text in the corpus was created with the spelling normalized according to a published orthographic standard for Plautdietsch. Each orthographically normalized token in these separate versions was cross-referenced with the token or tokens to which it corresponds in the original text via a unique identifier. Thus, the corpus maintains both the original, authorial spelling of each text, as was required to respect the orthographic wishes of each author, as well as a standardized representation of the same, with both versions available for later use in linguistic inquiry.
2. *Adaptation of an existing tagset to Plautdietsch*. Rather than attempt to develop an entirely new tagset for Plautdietsch, an existing set of conventions proposed for the assignment of part-of-speech tags to Standard German, the Münster Tagset German (MT/D; Steiner, 2001) was adapted to suit Plautdietsch. Where existing grammatical descriptions were in agreement, categories in the Standard German tagset, which are not found in Plautdietsch (e.g., a distinct morphological verb form representing the subjunctive aspect, which has merged with the simple past in Plautdietsch), were eliminated. Where grammatical descriptions of Plautdietsch were in disagreement, the more detailed categories of the larger tagset were generally preserved. This process resulted in a significant reduction in the overall number of categories for annotation, leaving 99 distinct part-of-speech tags. Adapting an existing set of published tagging conventions for a related language, while clearly not an option available to all minority languages, proved to be of benefit here, allowing greater attention to be given to those particular cases in which conventions of the source tagset appeared out of step with features of the

target language than might have been otherwise possible, had a comparable tagset for Plautdietsch been developed *de novo*.

3. *Probabilistic tagging*. In order to apply the adapted tagset to the now orthographically normalized Plautdietsch texts, corpus construction employed a language-independent, pure probabilistic tagger, Qtag (Tufis and Mason, 1998). Several features of Qtag motivated its selection over other, comparable probabilistic tagging systems, not the least of which was its provision of a reference implementation of the probabilistic tagging algorithm on which it relies. As this implementation was written in Java and made freely available for non-commercial use, and provided a well-documented application programming interface (API) supporting Unicode, it was anticipated that Qtag might be integrated into the current project with minimal expenditure of resources, financial and otherwise. Since the basic Qtag algorithm has been published in Tufis and Mason (1998), it would also had been possible to reimplement this system independently in another programming language or environment, had the need arisen. While Qtag is certainly not alone in the class of probabilistic taggers offering similar features, it nevertheless presents a reasonable point of departure into probabilistic tagging.

Of these three stages of corpus construction, the final one, in which the adopted tagset was applied to the normalized texts which comprised the corpus, proved to be the most involved, even with the assistance provided by a probabilistic tagger. As Qtag requires a set of correct tag-token pairs from which to induce its initial probabilistic model, it was not possible to apply the tagger to the entire corpus immediately, and no other corpora of Plautdietsch were available from which such training data could be drawn. Instead, corpus texts were tagged with the adopted tagset incrementally in an iterative, interactive process. At the beginning of this process, each corpus text was divided into c segments, or chunks, of n tokens. For the first text in the corpus, tags were assigned manually to each of the n tokens appearing in its first chunk, producing a total of n-correct tag-token pairs. Qtag was then trained on these first n tag-token pairs, forming an initial probabilistic model of the language, consisting of both a matrix of transition probabilities between the observed sequences of tags as well as a probabilistic lexicon of observed token-tag associations. This model was then used as input to the first iteration of tagging: taking this model as an indication of how tags are meant to be applied, Qtag was made to assign tags probabilistically to the n tokens of the next chunk of text. These probabilistic assignments were then corrected manually—the interactive portion of this process—and Qtag was retrained on the collection of $2n$-verified tag-token pairs, which were then available as training data. The same process was repeated for all remaining chunks in all remaining documents, having Qtag assign tags to each chunk on the basis of the corrected examples it had been trained on thus far, and these assignments then being corrected by hand and fed into the ever-growing set

of training data from which Qtag inductively built its model of lexical and tag-sequence probabilities.

This iterative, interactive process was repeated until all chunks in the corpus had been successfully tagged and corrected. This process thus made use of the ability of Qtag to assign tags probabilistically given even a small amount of manually corrected input, and to develop progressively more informed (and, with any luck, more accurate) models of the language and the tagset under consideration as greater amounts of corrected training data became available over subsequent iterations. All told, this corpus construction process, when applied to the present corpus, resulted in approximately 120,000 tokens of orthographically normalized Plautdietsch text, tagged entirely according to the adapted tagset.

3. Modeling corpus construction

While ultimately successful, as noted in the previous section, this iterative, interactive process of tag assignment proved to be the single most time-consuming and labor-intensive segment of the larger corpus construction procedure. As such, it was also the most crucial element to the completion or abandonment of corpus development plans. Had this stage taken more time to finish than was anticipated or more resources than had been allotted to it, tag assignment may have needed to be set aside or the entire corpus development plan reconsidered. The question might therefore be asked: what could have been done to reduce the burden of corpus construction as a whole, without lessening the quality of the resulting data? Was it necessary, for example, to provide Qtag with normalized spellings in advance of tagging? Would omitting this step, itself a considerable investment of time, have reduced to any significant extent the rate at which accuracy developed in the later iterative, interactive tagging process? Should the tagset adopted for application to the corpus texts have been more or less elaborate than it was? Should greater numbers of tokens have been tagged in each iteration of the tagging process?

It is clear that the modest success achieved in the present method of corpus construction in arriving at an application of the given tagset to the corpus data cannot reasonably be taken to imply anything more than that it was possible, given the noted investment of time and effort. Open questions remain as to whether or not better solutions to this same problem might have been found, thus decreasing overall resource expenditure while achieving the same final product—or indeed, which if any of the decisions made in planning corpus development may have borne most heavily upon the investment of effort ultimately required. The answers to these questions, however, are of immediate interest, to continued corpus development within the present project and potentially to comparable minority-language corpus construction projects as well and, thus, arguably deserve further attention.

In order to begin to address questions such as these, the present study opts to conduct computational simulations of different models of interactive, iterative

tagging. In these simulations, different combinations of parameters to the tagging process are varied systematically to represent distinct combinations of choices that might have been made during corpus planning. Having the final, corrected corpus on hand, it is possible to recreate in each automated simulation the iterative, interactive tagging process for the combination of parameters under investigation. That is, having on hand the final set of corrected training data for the entire corpus, it is possible to automate the training of Qtag on successively larger portions of these data, monitoring at each stage the accuracy of its tag assignments to the next chunk of the remaining corpus data. Simulations are thus able to observe and quantify the performance of a given constellation of corpus design decisions, relative to assumed goals and requirements; they can, in effect, undertake a simple exploration of the parameter space of possible corpus design decisions, comparing the relative merits of each such option according to the criteria used for evaluation.[1]

As parameters to the simulations conducted here, three classes of corpus design decisions are considered.

1. *Orthographic normalization*. Simulations of probabilistic tagging are performed using both orthographically normalized and orthographically unnormalized data.
2. *Chunk size*. Individual simulations vary the number of tokens to which tags are assigned and subsequently corrected in each round of iterative, interactive tagging. For the purposes of this study, 14 chunk sizes are considered (i.e., 100; 200; 300; 400; 500; 750; 1,000; 1,500; 2,000; 3,000; 4,000; 5,000; 7,500; and 10,000 tokens per chunk). Although the selection of token sizes made here is somewhat arbitrary, its range is arguably not altogether unrealistic for manual tag assignment correction. However, nothing prevents the consideration of other chunk sizes as well.
3. *Choice of tagset*. Simulations vary the complexity of the tagset being applied. Two new tagsets are defined as surjections from the categories of the original 99-tag tagset to sets of categories having 50 and 13 tags, respectively. This results in three tagsets of differing complexity being compared here: where the 99-tag tagset assigns distinct labels to the possible combinations of tense, person, and number features on inflected verbs, for instance, these labels all reduce to the single category of *V* (verb) in the 13-tag tagset.

Each combination of these parameters—each model of iterative, interactive tagging—is simulated and evaluated in terms of two measures. The first is the rate of accuracy development over time. Of interest here are models in which initial accuracy is high and increases rapidly as more training data are supplied. The second measure is the estimated amount of time required to produce each model. The total time required to apply tagset *t* to a given model *M* is estimated as a function of the time required for the initial, manual tagging of

the first chunk c_1 and the subsequent correction of automatic tag assignments of varying levels of accuracy for the remaining *c - 1* chunks:

$$(1) \qquad time_{total_t} = time_{manual_t}(c_1) + \sum_{i=2}^{C} time_{correction_t}(c_i, accuracy_t(c_i))$$

Estimates of the actual time requirement of manual tagging and correction at various error levels were gained through timed samples of these activities with each of the defined tagsets. While both of these metrics might be further refined and additional measures proposed by which to evaluate each simulation, they nevertheless provide in their present forms relatively intuitive means of assessing the models under consideration here.

4. Evaluating models of corpus annotation

For the purpose of exposition, the simulated models of part-of-speech tagging are divided into three classes, each concentrating upon one of the parameters introduced in the preceding section. We consider the effect of each level of these parameters upon the rate at which accuracy develops and the estimated time requirement observed for the simulated models while holding the levels of all other parameters constant, allowing us to compare the effects of each parameter individually. This procedure may be pursued exhaustively, exploring all possible combinations of parameter levels, or selectively, considering only those parameter combinations which are considered particularly promising by a given heuristic. The objection might reasonably be raised that the latter methodology, if applied blindly, may inadvertently obscure potential interactions between parameters—a particular combination of parameter levels may perform especially well or exceptionally poorly, and this important interaction be overlooked if the values of all other parameters except the one of interest are held constant. This is not an issue in the present simulations, where all parameter-level combinations have been simulated exhaustively and no such significant interactions noted. The present description of the results of simulation is therefore intended to reproduce the effect of each individual parameter, without suggesting that interactions between parameters might be safely disregarded in all such cases. We begin by considering the effects of orthographic normalization upon the rate of accuracy development and estimated time expenditure across simulations, followed by the same for chunk size and tagset choice.

5. Evaluating orthographic normalization

Guiding the investigation pursued in this section is the question raised previously: does orthographic normalization matter, either for the rate at which tagging accuracy develops over successive iterations or for the estimated overall time expenditure? The method of simulation adopted here offers one means of

addressing this question. Holding the choices of tagset and chunk size constant, we compare the results of simulations of tagging normalized and unnormalized data, thus isolating insofar as possible the effects of normalization, independent of the remaining decisions made in corpus planning.

Figure 1 and Figure 2 present the differences noted in accuracy development rates and estimated time requirements for the tagging of normalized and unnormalized data. A Wilcoxon signed rank test confirms a statistically significant difference between the rates of accuracy depicted in Figure 1 (n = 1,237, W = 742,550, $p < 0.0001$). While these figures concentrate upon the results of simulation for POS-99 (the original 99-tag tagset), similar, albeit less dramatic, relationships hold for POS-50 and POS-13 as well. On average, accuracy development rates are 20% lower for unnormalized data than for normalized data across all tagsets. As might be expected, these decreased accuracy rates correlate with increased estimated time requirements: the time required to tag the entire unnormalized corpus is estimated on the basis of these simulations to take on average 26 hours longer for POS-99, 15 hours longer for POS-50, and 11 hours longer for POS-13 than it would to tag the same corpus in orthographically normalized form. This suggests a considerable difference in both the rate of accuracy development and overall time requirements between the tagging of normalized and unnormalized text in this corpus. It would seem that, in the present corpus at least, orthographic normalization has a substantial effect upon both probabilistic tagging accuracy and estimated total time requirements—an effect observed across all tagsets and chunk sizes considered here.

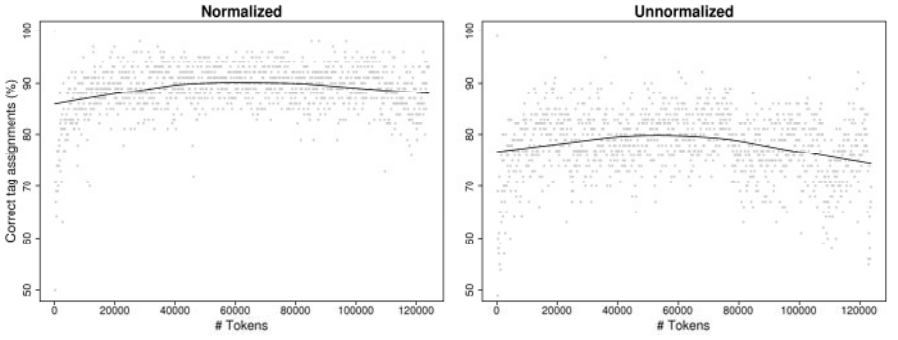


Figure 1. Rate of accuracy development observed in simulations with normalized and unnormalized data (POS-99, chunk size 100), with curves fitted by Lowess smoothers.

These results in turn raise an interesting question: what aspects of orthographically unnormalized text pose the greatest problems for the adopted methods of probabilistic tagging? While a thorough investigation of this question falls largely outside of the scope of the present investigation, it might be hypothesized that the increased number of orthographic variants found in the unnormalized corpus, serving to inflate the number of unique word forms (i.e., types) with which the probabilistic tagger must come to terms, in part causes

overall tagging accuracy to decrease. A simple negative correlation between type count and tagging accuracy would seem unlikely, however. Rather, the frequency of occurrence and consistency of tagging of individual types within the corpus, as well as the predictability of the part-of-speech categories of these types from the contexts in which they appear, may also be relevant to tagging accuracy, rendering this a less-than-trivial problem to explore, although one of potential relevance to corpus construction. Importantly, these hypothesized predictors of tagging accuracy are themselves quantitative measures derivable from corpora and tagging procedures, and thus open questions such as this to further quantitative investigation.

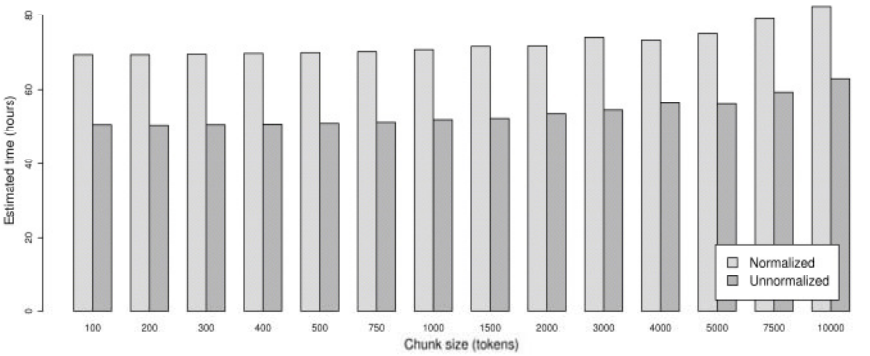


Figure 2. Estimated time expenditure for tagging normalized and unnormalized data across all specified chunk sizes (POS-99).

5.1 Evaluating chunk size

A procedure similar to the one used to assess orthographic normalization in the preceding section is adopted here to evaluate the effects of chunk size on the development of accuracy rates in tagging and the estimated overall time requirement of corpus construction. In this case, holding both the choice of tagset and orthographic normalization constant, simulations of tagging using different chunk sizes are conducted and their results compared. These results are presented graphically in Figure 3 and Figure 4. Since many chunk sizes result in similar accuracy rate measures, these rates are presented as histograms with three points for each chunk size (i.e., tagging accuracy rates at the beginning, middle, and end of tagging the corpus).

Inspection of these figures suggests an immediate difference between the relationship of chunk size and that of normalization to the assumed measures of accuracy development and estimated time investment. First, statistical investigation using Pearson's product moment correlation test suggests a general correlation between the rate of accuracy development and chunk size in initial (r = 0.6398, df = 12, p = 0.01373) and final (r = 0.9451, df = 12, p < 0.000001) stages of tagging, albeit not significantly in medial stages (r = -0.4887, df = 12, p

= 0.07617). Likewise, positive correlations between estimated time requirement and chunk size are noted for POS-99 (r = 0.9939, df = 12, p < 0.000001), POS-50 (r = 0.9970, df = 12, p < 0.000001), and POS-13 (r = 0.9976, df = 12, p < 0.000001).

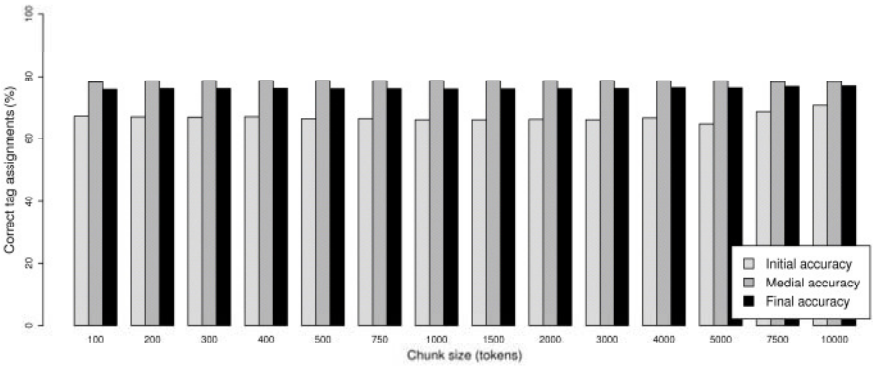


Figure 3. Rates of accuracy development observed in the initial, medial, and final stages of simulations of tagging across all specified chunk sizes (POS-99, normalized data).

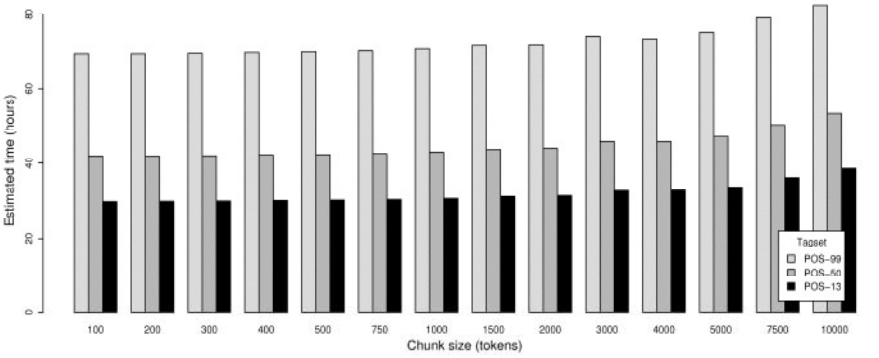


Figure 4. Estimated time expenditure of tagging in each tagset across all specified chunk sizes (normalized data).

While accuracy rates remain essentially the same for all chunk sizes less than or equal to 5,000 tokens, considerable differences are found in estimated time requirements, with smaller chunk sizes (roughly, those of less than 2,000 tokens) consistently taking less time than larger ones. The reason for this difference lies in the amount of time required to assign tags manually to the first chunk of corpus text: without the aid of automatically assigned tags (even incorrect ones), this stage of iterative, interactive tagging typically takes longer than later stages of correction. As the size of the first chunk increases, so too does the amount of time required to assign tags to each token in that chunk by hand, thus gradually coming to outweigh any potential benefits to overall accuracy that

might have accompanied providing the tagger with a greater amount of initial training data. In the present corpus, then, it would appear sensible to attempt to minimize the amount of time required to tag the first chunk by hand and to concentrate instead upon the correction of probabilistically assigned tags in the remainder of the corpus.

5.2 Evaluating tagset choice

To evaluate the contribution of tagset choice to the rate of accuracy development and estimated overall time expenditure, all other parameters are once again held constant, and the results of simulations of tagging with each of the three proposed tagsets are compared. These results are presented in summarized form in Figure 5.

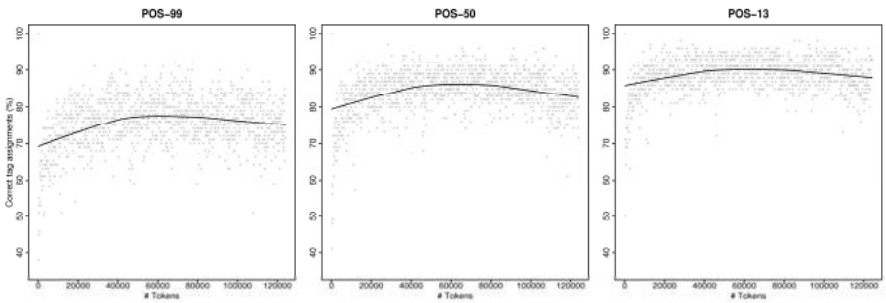


Figure 5. Rates of accuracy development across tagsets (normalized data, chunk size 100), with curves fitted by Lowess smoothers.

Much as was the case with the figures presented in preceding sections, Figure 5 suggests a clear difference between the three tagsets and brings to the fore a general negative correlation between tagset size and mean tagging accuracy which achieves near statistical significance here (r = -0.9956, df = 1, p = 0.0595). In their mean rates of accuracy development, an average 15% improvement is found for the minimal tagset, POS-13, over the full tagset, POS-99, regardless of whether or not the texts being tagged were normalized. Likewise, the estimated time requirement for applying POS-99 to normalized data, 80.5 hours, is more than double the 36.5 hours estimated to be needed to apply POS-13 to the same data. The choice of tagsets would thus appear to represent an important factor in the overall investment of effort required to achieve full tagging, even for a corpus of this size.

While these results are intriguing and may be of use in the present corpus development project, they should perhaps be interpreted with a degree of circumspection. The alternative tagsets considered here are direct adaptations of the full 99-tag tagset modified for use with Plautdietsch. It may be the case, however, that other, more varied tagsets may have fared better or worse when applied to the same corpus data, although this cannot easily be tested without some means of applying these tagsets to at least a sample of the present corpus

for comparison. Although the relatively simpler tagsets consistently achieved lower estimated time requirements and higher rates of accuracy development in simulations, the larger question nevertheless remains: what features of these simpler tagsets, beyond the restricted range of those labels they provide, might be cited to explain their respective degrees of success in tagging these data? What features of these or other tagsets make them more or less well-suited to the data and the probabilistic tagging system at hand? One might expect some degree of increase in accuracy by chance alone: all other things being equal, a smaller tagset provides fewer opportunities than a larger one for a probabilistic tagger to guess incorrectly, and thus might be anticipated to deliver, on the whole, more accurate responses. As in the case of orthographic normalization, however, a thorough answer to questions such as these pertaining to tagset design lies outside the purview of this study, which is concerned primarily with tagset application. Nevertheless, methods of evaluation such as the simulation-based techniques employed here might be of some use in developing quantitative measures of relative tagset complexity, and thus be helpful in addressing these open issues in corpus design as well.

6. Summary and conclusions

In the present case, evaluation of the results of simulation would appear to suggest the following guidelines as relevant to successful tagging:

1. With regard to orthographic normalization, improvements in the rate of tagging accuracy development may be substantial when working with normalized data (here, on the order of 20% more accurate). However, these gains must be weighed against the cost of normalization itself. Particularly in the case of minority languages, where no single orthographic standard (or appropriate spell-checking software, for that matter, if the process of orthographic normalization is to be partially automated) may necessarily exist, achieving consistent normalization may itself represent a considerable investment of effort.
2. When deciding on the size of chunks of corpus text to process in each iteration, smaller chunk sizes should be favored over larger ones. This recommendation is motivated by the observation that manual tagging may, in many cases, prove more time-consuming than correction of automatic tag assignments, regardless of the latter's accuracy. Relying on the probabilistic tagger early in corpus development to perform as much tag assignment as possible limits the amount of manual tagging required.
3. Less elaborate tagsets should be favored wherever corpus goals permit. While substantial gains in accuracy development rates and decreases in estimated time requirements were noted with the less detailed tagsets considered here, this observation should be interpreted with care, since no quantitative measures of tagset complexity have been established here.

> Any decisions to change the adopted tagset should be evaluated with attention to the requirements of the anticipated uses and users of the corpus: there is little benefit in applying a tagset which is too simple to be of use to those working with the finished corpus.

Such suggestions, however, must ultimately be measured not only against the quantitative estimates of accuracy development rates and overall time investment, but also against the qualitative requirements, available resources, and stated goals of the given corpus project. In this instance, where verbal complementation represents a primary focus of research, detailed coding for the inflectional features of Plautdietsch verbal morphology is needed. The choice of tagsets, then, is constrained to some extent by this requirement, although the cost of adopting a more complex tagset can be mitigated in part through orthographic normalization and the selection of a small chunk size, as the preceding simulations have suggested. Likewise, while a simpler tagset may have rendered it technically feasible to process orthographically unnormalized data, and thus avoid investment of resources in orthographic normalization, the goal of supporting exhaustive searches of the resulting corpus motivated this additional expenditure of time and effort. In short, quantitative measures cannot afford to be the sole means of assessing the relative merits of alternative corpus development strategies, although their application may indeed be of considerable benefit in corpus planning. Such measures form one important aspect of informed corpus development which exists within a larger rubric of goals and requirements. When taken together, these factors may suggest paths by which corpus construction can be guided to satisfactory completion from both quantitative and qualitative perspectives.

It is readily conceded here that determining the interactions of all such factors, whether quantitative or qualitative, in their relationship to tagging accuracy is likely impossible during corpus planning. Careful planning may well be able to anticipate many of the factors and their interactions relevant to the completion of corpus development, but the goals, requirements, and resources available to corpus construction are likely to change with the corpus as development proceeds. It is maintained here, however, that such planning and evaluation might still profitably enter into corpus construction as a regular part of the larger development process and to no less effect in the construction of minority language corpora. In the initial phases of corpus planning, consideration of reported results and guidelines proposed on the basis of other corpus construction projects, such as those put forward in this case study, might inform corpus development and suggest potential pitfalls which should be avoided. Introducing periodic evaluation, whether using simulation or other means, as an additional part of the iterative tagging process may further serve to identify problems during corpus development and present opportunities to make midstream changes as necessary. Even if all relevant interactions are not apparent in advance of corpus implementation, this would not seem to imply that corpus

development cannot benefit from previous experience in corpus design or from regular evaluation during corpus construction.

The selection of pure probabilistic methods over other methods available for tag assignment is also far from given. This decision, too, might be informed by consideration not only of the technical requirements of corpus development and the requirements implied by corpus end-use goals, but also the typological features of the language and the characteristics of the available sources of data, as noted previously. If one of the goals of corpus development is to integrate corpus documents with other available sources of linguistic data (e.g., existing digital dictionaries, word lists, collections of morphological parses, etc.), this may encourage the use of hybrid tools which permit concurrent lemmatization or other annotation. Whereas the present task appears well suited to the use of Qtag, benefitting from this tool's availability and simple integration into larger projects, this should not be taken to suggest that other, comparable tools might not be appropriate in the same or similar contexts or that their application to corpus development might not benefit from the processes of evaluation discussed here as well.

Computer-assisted methods of part-of-speech assignment present a range of complex technical and practical problems for the construction of modern corpora. As a stage of corpus development during which considerable resources must commonly be invested, corpus tagging represents a particular challenge for minority-language corpus development, where such resources are often limited, and thus an area in which quantitative, computational methods of evaluation might be of use in corpus development planning. While computational methods of tagging and evaluation are of clear importance to the progress of many corpus development projects, and thus arguably merit the attention which has been given to them here, it has been insisted here they arguably cannot afford to be the sole object of inquiry. Rather, consideration is also required of the larger context in which such methods are applied and of the resources, research requirements, and (socio)linguistic conditions which bear upon corpus construction as a whole.

Case studies of minority-language corpus construction present one means of contributing to an understanding of such problems in context. The results of such case studies, which are in most instances language and corpus-specific, might serve to offer general direction for further quantitative studies of corpus and tagset design, as several sections of this study have suggested. By the same token, case studies might offer an honest assessment of the challenges facing corpus construction and corpus-based language documentation in the use of contemporary computational techniques, providing guidelines from which similar projects might benefit. In bringing attention to practical issues encountered in the application of current computational methods to data from underrepresented languages, evaluations of minority-language corpus construction might thus serve a twofold purpose—at once presenting computational-linguistic research with additional real-world benchmarks by which their success under varied linguistic and sociolinguistic conditions might be assessed, while fostering through the

description of current corpus construction techniques the continued development of annotated corpora for a greater number of the world's languages.

7. Notes

1. Since the number of parameters to these models of corpus construction is limited in this case, the combinatorial space which these options form can be explored exhaustively without significant difficulty (in part because simulations may be conducted in parallel, being computationally independent of one another) and all possible models thus compared with an even hand. This may not always be the case, however, since the number of possible models increases essentially exponentially in the number of parameters under consideration. With parameter-rich simulations, then, other methods of estimating the relationships of individual parameters to measures of interest may be required.

References

Church, K. W. (1988), "A stochastic parts program and noun phrase parser for unrestricted text," in: *Proceedings of the 2nd conference on applied natural language processing*. ACM, 136–143. Online at: <http://www.aclweb.org/anthology-new/A/A88/A88-1019.pdf>.

DeRose, S. J. (1988), "Grammatical category disambiguation by statistical optimization," *Computational linguistics*, 14(1): 31–39.

Epp, R. (1993), *The history of Low German and Plautdietsch: tracing a language across the globe*. Hillsboro, KS: The Reader's Press.

Epp, R. (1996), *The spelling of Low German and Plautdietsch*. Hillsboro, KS: The Reader's Press.

Epp, R. (2002), *De Jeschicht von Plautdietsch* [The history of Plautdietsch]. Kelowna, BC: unpublished m.s.

Gordon, R. G., Jr. (2005), "Plautdietsch," *Ethnologue: languages of the world, 15th edition*. Dallas, TX: SIL International. Online at: <http://www.ethnologue.com>.

Hooge, D. J. (1973), "Das Verb in der Parataxe und Hypotaxe statistisch gesehen [The verb in parataxis and hypotaxis, from a statistical perspective]," *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, 26: 328–341.

Klassen, H. N. (1969), "Untersuchungen zum grammatischen Bau der niederdeutschen Mundart im Gebiet Orenburg (RSFSR) [Investigations of the grammatical structure of the Low German dialect in the region of Orenburg (USSR)]," *Wissenschaftliche Zeitschrift der Martin-Luther-Universität Halle-Wittenberg, Gesellschafts- und Sprachwissenschaftliche Reihe*, 18(6): 27–48.

Klassen, H. N. (1993), *Mundart und plautdietsche Jeschichte. Ut dem Orenburgschen en ut dem Memritjschen (Rußland)* [Dialect and Plautdietsch stories. From Orenburg and Memrik (Russia)]. Marburg: N. G. Elwert Verlag.

Loewen, H.. and A. Reimer (1985), "Origins and literary development of Canadian-Mennonite Low German," *Mennonite quarterly review*, 59: 179–186.

McEnery, T., and N. Ostler. (2000), "A new agenda for corpus linguistics—working with all of the world's languages," *Literary and linguistic computing*, 15(4): 403–420.

Nieuweboer, R. (1999), *The Altai dialect of Plautdiitsch: West-Siberian Mennonite Low German*. Munich/Newcastle: Lincom Europa.

Steiner, P. (2003), *Das revidierte Münsteraner Tagset Deutsch (MT/D). Beschreibung, Anwendung, Beispiele und Problemfälle* [The revised Münster Tagset for German (MT/D). description, application, examples, and problematic cases]. Online version: <http://xlex.uni-muenster.de/Portal/MTPD/tagsetDescriptionDE.ps>.

Tufis, D., and O. Mason (1998), "Tagging Romanian texts: a case study for QTAG, a language independent probabilistic tagger," in: *Proceedings of the 1st international conference on language resources and evaluation*. ELRA, 589–596. Online at: <http://www.racai.ro/~tufis/papers/Tufis-Mason-LREC1998.pdf>.

Exploring a corpus of scientific texts using data mining

Elke Teich

Technische Universität Darmstadt, Germany

*Peter Fankhauser**

L3S, Leibniz Universität Hannover, Germany

Abstract

We report on a project investigating the linguistic properties of English scientific texts on the basis of a corpus of journal articles from nine academic disciplines. The goal of the project is to gain insights on registers emerging at the boundaries of computer science and some other discipline (e.g., bioinformatics, computational linguistics, computational engineering). The questions we focus on in this paper are (a) how characteristic is the corpus of the meta-register it represents, and (b) how different/similar are the subcorpora in terms of the more specific registers they instantiate? We analyze the corpus using several data-mining techniques, including feature ranking, clustering, and classification, to see how the subcorpora group in terms of selected linguistic features. The results show that our corpus is well distinguished in terms of the meta-register of scientific writing; also, we find interesting distinctive features for the subcorpora as indicators of register diversification. Apart from presenting the results of our analyses, we will also reflect upon and assess the use of data mining for the tasks of corpus exploration and analysis.

1. Introduction

The broader context in which the present paper is placed is corpus comparison. Corpus comparison is involved in many areas of corpus linguistics, ranging from the comparative analysis of registers/genres, varieties, and languages (including translations), from both a synchronic and a diachronic perspective (Biber, 1988, 1995; Mair, 2006, 2009; Teich, 2003). With this comes a concern for methodologies of corpus comparison, laying out the principal work flows in corpus compilation and corpus processing (annotation, query). In this context, questions of data analysis have recently received some attention, addressing the important issue of appropriate statistical measures for interpreting quantitative data. Examples are Kilgariff (2001), who discusses a range of statistical techniques and their applicability to lexically based corpus comparison, and Gries (2006), who presents a method of measuring variability within and between corpora.

The primary concern of the present paper is a methodological one. The concrete background of our work is a research project on the specifics of language use in interdisciplinary scientific contexts, with a focus on scientific

registers at the boundaries of computer science (such as bioinformatics, computational linguistics or computational engineering). The research questions we are interested in include the following: What are the linguistic effects of a scientific discipline coming into contact or merging with computer science? To what extent are the linguistic conventions of the original discipline retained? Are there any tendencies to adopt the language of computer science? Or are there new registers developing?

The data we work on is the *Darmstadt Scientific Text Corpus* (DaSciTex), which contains full English scientific journal articles compiled from 23 sources and covering nine scientific disciplines. The corpus comes in two versions: a large one comprising around 19 million words, and a small one comprising around one million words (Holtz and Teich, in preparation).[1] The corpus includes texts from the broader areas of humanities, science, and engineering and has a three-way partition (see also Figure 1 for a diagrammatic representation):

A computer science
B mixed disciplines
- **B1** computational linguistics
- **B2** bioinformatics
- **B3** computer-aided design/construction in mechanical engineering
- **B4** microelectronics/VLSI

C pure disciplines
- **C1** linguistics
- **C2** biology
- **C3** mechanical engineering
- **C4** electrical engineering

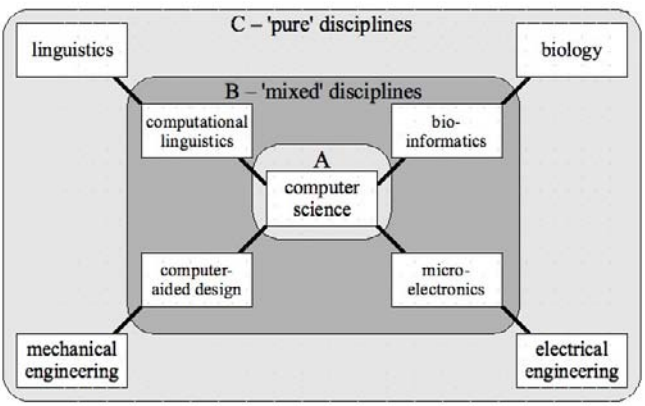


Figure 1. Design of DaSciTex.

At the methodological level, the project is concerned with developing a methodology for corpus comparison with a special view to fine-grained linguistic variation: What are the most distinctive features between the corpora under investigation and how can we obtain these features? To approach this question,

we explore a set of techniques from an area known as data mining (see Chakrabarti [2003] and Witten and Eibe [2005] for comprehensive introductions, as well as Manning and Schütze [1999: chapter 16] and Sebastiani [2002] for overviews). To our knowledge, except for clustering (e.g., Biber, 1993), data-mining methods have hitherto hardly been explored in the context of corpus comparison.

We have carried out several analyses using data mining that address the following two types of questions:

1. How well is the corpus distinguished in terms of features characterizing the meta-register of scientific writing?
2. How different/similar are the subcorpora in terms of features characterizing the individual registers?

To explore the first question, we have compared the DaSciTex corpus with the registerially mixed FLOB corpus.[2] If DaSciTex represents scientific writing, then the texts contained should exhibit some typical properties of this meta-register, such as (relative) abstractness, technicality, and informational density (see Halliday [1985]; Halliday and Martin [1993]), which are not exhibited by a registerially mixed corpus such as FLOB. To explore the second question, we have compared the subcorpora of DaSciTex in order to determine the relative position of the mixed disciplines vis à vis their corresponding pure disciplines and computer science. Inspecting the texts in the corpus, we observed that one potential source of difference lies in the roles and attitudes participants adopt in the discourse situation (see Section 2 on parameters of register variation). One indicator of this is the self-construal of the authors in terms of the types of activities engaged in. For some examples see (1)–(3) below, (1) instantiating two material processes with *we* as Actor, (2) a mental process with *we* as Senser, and (3) a verbal process with *we* as Sayer.

(1) *We analyze* and *compare* different queue policies.…
(2) *We believe* that competitive analysis gives important insights.…
(3) *We argue* that novel instances of verb adjective sequences are based on analogies to previous experiences.…

Analyses of this kind, testing for a variety of potential register features, will bring out the differences and commonalities between the registers under investigation.

The remainder of the paper is organized as follows: we start with an introduction to the underlying linguistic-theoretical framework we work with, Systemic Functional Linguistics (SFL) (Halliday, 2004), which has at its core a model of register variation (Section 2). This is followed by the presentation of the analyses we have carried out using selected techniques of data mining (Section 3). Section 4 concludes the paper with a summary and discussion.

2. Theory and model of register variation: Systemic Functional Linguistics

In descriptive linguistics, the notion of register refers to linguistic variation according to *use in context* (see Quirk et al., 1985; Biber et al., 1999). Register variation is a well-researched topic that typically requires working in a corpus-based fashion. In order to account for register differences in a corpus of texts, one needs a sound model of register variation that allows addressing the research questions involved. One such model is Systemic Functional Linguistics (SFL) (Halliday, 2004). The notion of register is at the core of the language model put forward by SFL (Halliday et al., 1964; Halliday, 1985; Halliday and Hasan, 1989; Matthiessen, 1993). SFL considers language a multi-dimensional resource for making meaning. The two dimensions of interest here are *stratification* and *instantiation* (see Figure 2 below). According to stratification, the linguistic system is organized along the levels of lexico-grammar, semantics, and context, where lexico-grammar is taken to realize semantics and semantics is taken to realize context. Instantiation refers to the relation between the linguistic system and a text (i.e., an instance). Each instance is characterized by the selection of particular linguistic (semantic, lexico-grammatical) features according to a context of situation. This situated language use results in *registers* or *text types*. For example, different sets of linguistic features will be chosen by speakers involved in a casual conversation compared to a highly structured and planned discourse, such as a written academic paper.

In SFL-based analysis of texts, an account of the contextual configuration forms an integral part. The instrument provided for accounting for a contextual configuration is given by three parameters that are said to characterize the level of context. These are *field*, *tenor*, and *mode* (see Quirk et al., 1985, who suggest a similar classification into field, attitude, and medium of discourse). Field of discourse is concerned with subject matter and the goal orientation of a text (e.g., expository, narrative, instructional). At the level of lexico-grammar, field is reflected in configurations of processes and the participants therein, such as actor, goal, medium, and accompanying circumstantials of time, place, manner, etc. (see again examples (1)–(3) in Section 1 above). Tenor of discourse is concerned with the roles and attitudes of the participants in a discourse. Linguistic reflexes can be found in choices of mood and modality as well as appraisal. Mode of discourse is concerned with the role language itself plays in the discourse (e.g., whether it is substantive or ancillary, whether the channel of communication is visual and/or auditive, whether the medium is written or spoken). Linguistic reflexes of this parameter are primarily textual (e.g., thematic structure, information structuring, informational density vs. grammatical intricacy). A register is thus constituted by particular settings of these three parameters—the contextual configuration—together with the sets of linguistic features *typically* chosen according to that contextual configuration (see Halliday and Hasan, 1989).

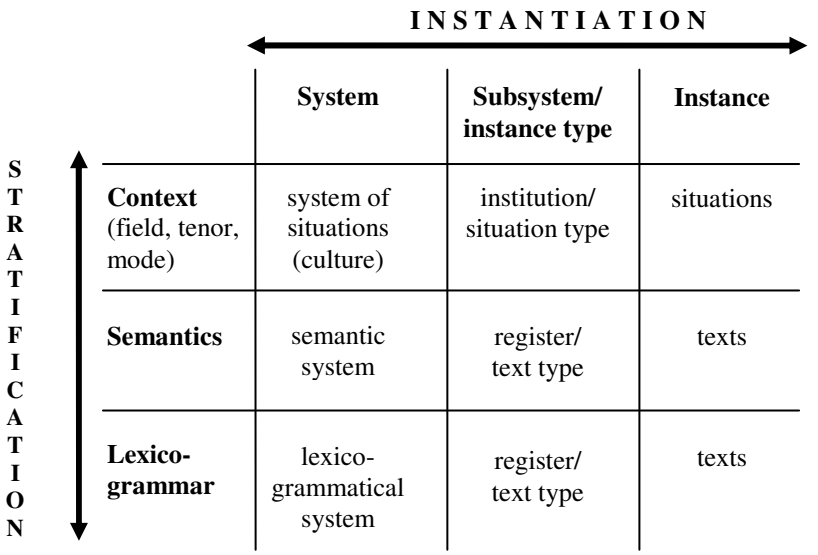


Figure 2. Register, stratification, and instantiation.

An analysis of the register(s) of a text or set of texts crucially involves statements about the distribution of features, i.e., it is a quantitative account (see Halliday, 2005). Also, since a text may exhibit some of the features typical of a register but not others (or, in other words, for a text to belong to a given register is a matter of degree), it is desirable to be able to bring out this kind of fuzziness in the analysis. Finally, since there is typically more than one feature involved in register differentiation, multivariate techniques qualify better than univariate ones (see again Biber, 1993). A set of methods that offer most of the desirable functionality is provided by an area known as data mining. The following section describes the analyses we have carried out on our corpus addressing the questions posed in Section 1, employing data-mining techniques such as feature ranking, clustering, and classification.

3. Data mining for corpus comparison: experimental setup and results

To approach the questions formulated in Section 1, we have carried out a number of analyses on the basis of selected, potentially discriminatory features using the WEKA data-mining platform (Witten and Eibe, 2005). In this section we describe the two setups for these analyses and their results. The first setup addresses the first question by comparing DaSciTex with FLOB (Section 3.1); the second addresses the second question, comparing the subcorpora in DaSciTex (Section 3.2).

3.1 Analysis (1): comparing DaSciTex with FLOB

The data sets we work with in these analyses are the small version of DaSciTex (around one million tokens, 186 texts) and FLOB (also containing around one million tokens, 405 texts). Both corpora have been tokenized and part-of-speech tagged with the Tree Tagger (Schmid, 1994).[3] This first set of analyses aims at comparing the DaSciTex corpus as an instance of scientific writing with the FLOB corpus, which contains instances of various different registers (including a partition of scientific writing).

We have investigated the following candidate features as potential indicators of scientific writing, focusing on the properties of abstractness, technicality, and informational density:

- the relative number of nouns (NN), lexical verbs (VV), and adverbs (ADV) as possible indicators of abstract language;
- the standardized type-token ratio (STTR) as a potential indicator of technical language;
- the average number of lexical words per clause (LEX/C), i.e., lexical density, as a measure for the informational density of a text.

Table 1 gives the overall averages for FLOB and DaSciTex for these features evaluated for their discriminatory force by the technique of Information Gain (IGain) and, for comparison, by the T-Test. Both IGain and T-Test measure how well a feature distinguishes between classes.[4] For a set of features these measures also provide a ranking: the features in the table appear ranked in the order of top (highest discriminatory force: STTR) to bottom (lowest discriminatory force: VV) according to IGain. Note that this ranking matches the ranking by the T-Test exactly.

Table 1. Results for selected features comparing DaSciTex and FLOB.

	FLOB	**DaSciTex**	**IGain**	**T-Test**
STTR	43.600	34.000	0.48	23.8
ADV	0.056	0.034	0.33	21.2
NN	0.270	0.330	0.33	-18.3
LEX/C	6.160	8.390	0.26	-15.6
VV	0.114	0.097	0.12	10.3

The best discriminator is the STTR, with texts in the DaSciTex corpus having a significantly lower type token ratio than the texts in FLOB.[5] The T-statistics is 23.8, which is well above the critical value of 1.9 for 0.95 confidence. Accordingly, the IGain of STTR is also fairly high with 0.48. The next two features are the relative numbers of adverbs and nouns. DaSciTex has a larger number of nouns and a smaller number of adverbs than FLOB, a difference that is again significant, as shown by both the T-Statistics and IGain. The average number of lexical words per clause is ranked as the fourth feature. Again, DaSciTex has a larger number of lexical words per clause than FLOB.

Interestingly, the relative number of lexical verbs, shown here as the last feature, is a less strong discriminator than the number of adverbs, but still the number of verbs is significantly smaller in DaSciTex than in FLOB.

We have also investigated a number of other candidate features, including the number of words per sentence (T-Statistics: 3.4) and the number of clauses per sentence (T-Statistics: -7.6) as possible alternatives for LEX/C, and the ratio of lexical words vs. function words. While all these features have a T-Statistics above the critical value, they are far less discriminatory than the five features given above.

IGain and T-Statistics measure how well an individual feature distinguishes between classes. In order to better understand how these features *together* distinguish DaSciTex and FLOB, we have used them to train a classifier based on a linear support vector machine and clustered them based on K-means. See Table 2 below for the results. The top four features achieve a classification accuracy of 90% (with ten-fold cross validation) and a clustering accuracy of 84%.[6] While the clustering accuracy is naturally lower than the classification accuracy, it shows that the features separate the two corpora into two clusters, with one cluster containing mainly texts from DaSciTex and the other one mainly texts from FLOB. Clearly, one cannot expect 100% accuracy, as FLOB is composed of several registers including official reports (H - Government) and scientific writing (J - Learned), which are expected to have similar characteristics as the texts in DaSciTex with respect to the investigated features. Indeed, most of the texts from H and J are grouped into the DaSciTex cluster. When these are removed from FLOB, leaving 297 texts, the classification accuracy goes up to 97%, and the clustering accuracy goes up to 92%, as shown in Table 2.[7] Here, the most discriminatory feature alone—STTR—achieves a classification accuracy of 91%.

Table 2. Results for classification and clustering comparing (a) DaSciTex and FLOB and (b) DaSciTex and FLOB minus H, J (FLOB').

	Classification	**Clustering**
DaSciTex vs. FLOB	90%	84%
DaSciTex vs. FLOB'	97%	92%

Not considering H and J will then also increase the results for IGain and T-Test, as shown in Table 3 below.

Table 3. Results for selected features comparing DaSciTex and FLOB'.

	FLOB'	**DaSciTex**	**IGain**	**T-Test**
STTR	45.3	34.0	0.75	29.5
ADV	0.060	0.034	0.50	23.8
NN	0.27	0.33	0.41	-19.0
LEX/C	5.76	8.39	0.38	-18.4
VV	0.12	0.097	0.19	12.2

Thus we can conclude that type-token ratio, number of nouns and adverbs, and number of lexical words per clause distinguish DaSciTex from FLOB quite well. However, within the subcorpora of DaSciTex, these features are not distinctive. For example, when contrasting computer science and the mixed disciplines (A+B1+B2+B3+B4) with all pure disciplines (C1+C2+C3+C4), the IGain of all features but STTR is 0, with STTR still a very low 0.09. This indicates that DaSciTex is not only well distinguished from FLOB, but also rather coherent in itself with respect to these features.

3.2 Analysis (2): comparing subcorpora in DaSciTex

The data set we work with in these analyses is the full set of texts in DaSciTex (1,843 texts with about 19 million tokens). The texts are tokenized and part-of-speech tagged with the Tree Tagger (Schmid, 1994). This second type of analysis aims at comparing the subcorpora in DaSciTex to see how well the registers are discriminated.

As pointed out above, shallow features such as a low type-token ratio clearly characterize the meta-register of scientific writing, but they cannot distinguish between individual disciplines. Of course one can expect that disciplines are well distinguished by their subject-specific vocabulary represented mainly by nouns and to a lesser extent by verbs. To analyze this, we have selected the 500 most distinctive nouns in terms of their IGain and transformed the texts to their term vectors representing the frequencies of these 500 nouns. This representation we have used to train and test a classifier that classifies texts into the nine disciplines.[8] The achieved classification accuracy is 96%, and the misclassifications, which indicate overlaps between the subcorpora, exhibit an interesting pattern. See the confusion matrix in Table 4, where each row gives the predicted classes for an actual class.

Table 4. Confusion matrix for classification by nouns.

C/P	A	B1	B2	B3	B4	C1	C2	C3	C4	Sum
A	**217**	**0**	**2**	**2**	**1**	0	0	0	**5**	227
B1	**3**	**76**	0	0	0	**10**	0	0	0	89
B2	**3**	1	**275**	0	0	0	**6**	0	0	285
B3	**3**	0	0	**215**	0	0	0	**2**	**4**	224
B4	**1**	0	0	1	**204**	0	0	0	**0**	206
C1	0	**4**	0	0	0	**95**	0	0	1	100
C2	0	0	**5**	0	0	1	**236**	0	0	242
C3	0	0	0	**0**	0	0	0	**246**	**6**	252
C4	**4**	1	0	**4**	**0**	1	0	**5**	**203**	218

A computer science
B1 computational linguistics; **B2** bioinformatics; **B3** computer-aided design; **B4** microelectronics
C1 linguistics; **C2** biology; **C3** mechanical engineering; **C4** electrical engineering

The main diagonal (in bold) gives the number of correctly classified texts. Twenty-eight misclassifications (1.5%) occur between C4 and other engineering disciplines (A, B3, C3) (shaded in dark grey). Twenty-seven misclassifications (1.5%) occur between a mixed discipline (B1 through B4) and its corresponding pure discipline (C1 through C4) (two secondary diagonals, shaded in light grey). Fifteen misclassifications (0.8%) occur between computer science (A) and one of the mixed disciplines (shaded in grey), and only six otherwise. Thus we can observe that the engineering disciplines have the largest overlap, and the mixed disciplines have a larger overlap with their corresponding pure disciplines than with computer science, but overall, the overlap is fairly small.

Classification with the top 250 lexical verbs still achieves a fairly high accuracy of 87% and confirms the general pattern of overlap (see Table 5). We get 97 misclassifications (5.3%) among the engineering disciplines (C4 vs. A, B3, C3), 66 misclassifications (3.6%) between a mixed discipline and its corresponding pure discipline, 28 misclassifications (1.5%) between computer science and mixed disciplines, and 48 misclassifications (2.6%) otherwise.

Table 5. Confusion matrix for classification by verbs.

C/P	A	B1	B2	B3	B4	C1	C2	C3	C4	Sum
A	**200**	**3**	**1**	**9**	**2**	0	1	1	**10**	227
B1	**5**	**65**	5	1	1	**11**	0	0	1	89
B2	**2**	7	**266**	2	1	1	**3**	1	2	285
B3	**5**	4	0	**180**	1	0	0	**17**	**17**	224
B4	**1**	2	0	2	**200**	0	0	0	**1**	206
C1	0	**9**	0	1	0	**90**	0	0	0	100
C2	0	0	**6**	1	0	2	**229**	4	0	242
C3	1	0	1	**13**	1	1	2	**222**	**11**	252
C4	**13**	0	1	**32**	**6**	0	0	**14**	**152**	218

A computer science
B1 computational linguistics; **B2** bioinformatics; **B3** computer-aided design; **B4** microelectronics
C1 linguistics; **C2** biology; **C3** mechanical engineering; **C4** electrical engineering

In order to further analyze how the usage of verbs differs across disciplines, we investigate the colligation patterns of verbs with the pronoun *we* in Subject position. As illustrated in Section 1, this is interesting from the point of view of tenor of discourse (self-construal of the authors). These patterns can be extracted fairly accurately from the part-of-speech tagged corpus by means of a regular expression that selects all lexical verbs following *we*, possibly interleaved with adverbs and auxiliary verbs. Moreover, we split the individual texts into chunks of 30 subsequent occurrences of *we* + verb to balance the different text lengths in DaSciTex. Again, we take only the top 250 verbs into account.

Table 6 gives the confusion matrix for the triple of A (computer science), B1 (computational linguistics), and C1 (linguistics), uniformly sampled, such that each register contributes the same number of instances (234). The achieved

classification accuracy is 81% (87% for the full set, which contains more instances for A). As is to be expected, this is lower than the accuracy achieved by classification with all verbs. Again, the largest number of misclassifications (79 = 11.3%) occurs between B1 and C1, followed by 40 misclassifications (5.7%) between A and B1. Only 6 instances from A are misclassified as C1 or vice versa.

Table 6. Confusion matrix for classification by *we* + verb.

C/P	**A**	**B1**	**C1**
A	**210**	**21**	3
B1	**19**	**168**	**47**
C1	3	**32**	**199**

A computer science; **B1** computational linguistics; **C1** linguistics

Table 7. The nine most typical *we* + verb for each pair of subcorpora.

A vs. B1	**A vs. C1**	**B1 vs. A**	**B1 vs. C1**	**C1 vs. A**	**C1 vs. B1**
show	define	train	describe	argued	turn
prove	use	adopt	collect	argue	speculate
present	show	describe	examine	turn	feel
choose	present	induce	simplified	don	coded
save	denote	examine	use	read	assume
obtain	save	constrain	separated	examine	met
touch	evaluate	combined	evaluated	feel	read
get	describe	downloaded	given	suggesting	find
proved	obtain	separated	define	saw	presenting

A computer science; **B1** computational linguistics; **C1** linguistics

Because we use a linear support vector machine for classification, which assigns a negative or positive weight for each feature, we can also determine the most typical verbs for each subcorpus.

Table 7 shows the nine most typical verbs for each pair of subcorpora. Verbs typical of A in contrast to B1 and C1 are shaded in light gray, verbs typical of B1 compared to A and C1 are shaded in (medium) gray, and verbs typical of C1 compared to A and B1 are shaded in dark gray. Also, from this perspective, B1 is clearly positioned between A and C1; *define*, *use*, *evaluate*, *describe* are typical of A and B1 in contrast to C1, and *examine* is typical of B1 and C1 in contrast to A. At a more abstract level, we can say that the types of activities authors typically engage in in computer science texts (A) are of a formal nature (*prove*, *define*), whereas the authors in computational linguistics texts (B1) act experimentally *(collect*, *examine*), and in linguistics (C1) they act verbally (*argue*) as well as cognitively (*see*, *feel*).

4. Summary and discussion

In this paper, we have explored selected data-mining techniques for the purpose of analyzing register discrimination. The overarching question we are interested in is whether, in a situation of register contact between scientific disciplines, something like "interdisciplinary language" is emerging. Addressing this question involves register comparison. At a more technical level, this is an exercise in corpus comparison. In order to carry out corpus comparisons, we have set up an infrastructure for corpus processing that includes some standard tools, such as sentence splitting, tokenization, part-of-speech tagging, etc. (see Teich, 2009, for a typical processing pipeline). This provides the basis for corpus query and data analysis employing methods of data mining.

The crucial issue in this endeavor is how to discover those linguistic features that are good indicators of register differences, provided they exist. We have explored a set of features potentially distinguishing between the meta-register of scientific writing and other registers, on the one hand, and between individual registers of scientific writing, on the other hand. In the first type of analysis, we have compared the DaSciTex corpus with the registerially mixed FLOB corpus with regard to the properties of abstractness, technicality, and informational density using the following features: part-of-speech distribution, type-token ratio, and number of lexical items per clause. To determine the discriminatory force of these features, we have employed three data-mining methods: feature ranking, clustering (an unsupervised machine-learning method), and classification (a supervised machine-learning method). The results are consistent on all three methods, providing evidence that DaSciTex is well distinguished from FLOB according to the selected features.

In the second type of analysis, we have compared the subcorpora in DaSciTex with regard to selected lexico-grammatical features (nouns, verbs, *we* + verb) using classification. Here, the results are also conclusive: the subcorpora are lexically well distinguished. While this would be more or less expected because nouns and verbs are the main carriers of subject matter (i.e., they realize field of discourse), what is interesting to observe are the misclassifications arising between the subcorpora. We have suggested that these misclassifications might be systematic, indicating a similarity between some subcorpora but not others. Here, the consistent pattern for all analyzed feature sets is that the mixed disciplines (B corpora), the pure disciplines (C corpora), and computer science (A corpus) are well distinguished from one another, while at the same time the mixed disciplines are more similar to their corresponding pure disciplines than to computer science, and the engineering disciplines exhibit the largest similarity. This tendency is corroborated by a number of other studies in which we looked at the grammatical preferences of selected parts of speech (e.g., noun/verb colligations) (Holtz and Teich, in preparation). These analyses were conducted using more traditional descriptive statistical techniques as well as similarity measures between corpora. The results point in the same direction: triples of A-B-C corpora are clearly distinct with regard to the features investigated according to

various measures (e.g., chi-square, cosine distance, classification) *and* the B corpora (mixed disciplines) are "closer" to the C corpora (pure disciplines) than to the A corpus (computer science). At a more abstract level, this means that register contact results in mixed registers that, however, cannot deny their points of origin. Finally, investigating *we* + verb, we have found some colligations specific to the mixed disciplines. Provided that more evidence can be found of such distinctive lexico-grammatical patterns, one could conclude that interdisciplinary registers negotiate between integration and identification: they integrate the linguistic conventions of two different disciplines, while at the same time attempting to create their own identity as a unique discipline.

To assess the proposed methodology, three types of comments are in place. First, concerning the application of data mining, while simple descriptive statistics, such as a statistical test on a feature distribution, fulfills a similar purpose, there is a two-fold added value in using data-mining methods. In addition to ranking features by their individual discriminatory power, we can explore their *collective* contribution to register discrimination (multivariate analysis). Also, having available information about misclassifications in the form of the confusion matrix, we can investigate the context of typical features/terms in correctly classified and in misclassified texts, analyzing differences and commonalities between registers at class level as well as at instance level. From the perspective of data mining, register analysis is an interesting application because we are not primarily concerned with finding the most discriminatory feature set for optimal classification but rather with analyzing patterns of misclassifications for understanding register diversification. Second, from the viewpoint of register analysis, the features we have investigated here are obviously rather shallow, operating with strings and parts of speech. As part of the methodology, these have to be related to contextual configurations in a more principled way (consider again Biber's [1993, 1995] work on interpreting feature bundles in terms of more abstract dimensions of register variation). While words exhibit a strong discriminatory force between registers (i.e., an analysis on the basis of words provides a good classification result), they do not offer much interpretation space other than in terms of *field* variation ("a text t1 *is about* x and a text t2 *is about* y") (see again Section 2). Lexis-in-context (e.g., word/part-of-speech bi-grams), on the other hand, may not be such a strong discriminator (i.e., there will be more misclassifications), but it offers more interesting interpretation directions ("a text t1 *construes* x as y," "a text t2 *construes* x as z"). This is critical to get a grasp of other parameters of register diversification, such as writer-reader role relations (*tenor* variation) or textual organization (*mode* variation).[9] Third, from the perspective of SFL, we have proposed a possible operationalization for modeling register variation that allows interpreting feature distributions in terms of their contributions to register diversification. Here, a crucial aspect is the opportunity of representing the inherent fuzziness of registers: any one text purportedly belonging to a particular register may be more or less exhibiting the linguistic properties typically ascribed to that register. This

can be read off the confusion matrix, which thus turns out to be a convenient instrument for human inspection of analysis results.

In our future work we plan to carry out more analyses on the basis of aggregated linguistic information (such as part-of-speech n-grams) in order to explore other parameters of variation. For example, in the area of mode of discourse, colligations of nouns at the level of the nominal group could be a source of interesting differences between registers. Also, we are currently annotating a part of the corpus for process types and Theme-Rheme structure (Schwarz et al., 2008). We plan to use these annotations as a basis for training a classifier to annotate larger amounts of data from the DaSciTex corpus in order to be able to analyze differences and commonalities between registers at higher levels of linguistic organization.

5. Notes

* We are grateful to *Deutsche Forschungsgemeinschaft* (DFG), who support this work as grant TE 198/1-1 *Linguistische Profile interdisziplinärer Register* (Linguistic Profiles of Interdisciplinary Registers). Many thanks also go to Richard Eckart and Monica Holtz for their collaboration in corpus compilation and basic processing of the corpus, as well as to the anonymous reviewer for making helpful suggestions for improving our paper.

1. The corpus was compiled from pdf files, which were automatically transformed into plain text. The resulting data is not completely clean (e.g., erroneous splitting/contraction of tokens). For some types of investigations, one can live with this quality of data, but for others it is crucial to have them absolutely clean. We thus decided to manually clean a one million–token extract from the corpus, which is referred to here as the "small version."
2. FLOB was chosen for two reasons: first, it was readily available for us to carry out our own processing (POS-tagging, parsing, manual annotation); second, it is comparable in size to the small version of DaSciTex.
3. The reported accuracy of the TreeTagger is 96%.
4. IGain measures the reduction of uncertainty about a class *C* (e.g., FLOB vs. DaSciTex) when knowing an attribute *A* (e.g., STTR); more formally, it is defined by $H(C) - H(C|A)$, with $H(C)$ the entropy of *C*, and $H(C|A)$ the conditional entropy of *C* given *A*. The T-Test tests whether the observed means of two (normally distributed) populations differ significantly. We employed Welch's T-Test, assuming unequal variances. Note that unlike IGain, T-Test takes into account sample size.
5. To avoid the effect of different text lengths on type-token ratio, we have employed the standardized TTR.
6. A linear support vector machine is a linear classifier that separates two classes with a maximum margin between the two classes. We have used

the standard SVM implementation shipped with Weka. Other classifiers, such as naïve Bayes and decision tree learners, achieve a similar accuracy. The clustering accuracy is lower, because clustering operates unsupervised whereas classification operates supervised.

7. Gries (2006) provides a more systematic account on analyzing the variability of a single feature in a hierarchical corpus organized into registers and subregisters.
8. In more detail, the term frequencies are measured by TF/IDF (term frequency by inverse document frequency). Classification is performed by means of ten-fold cross validation, i.e., the sample is systematically split ten times into 90% training data and 10% testing data, and accuracies are averaged over the ten runs. Both are fairly standard procedures in the field of text classification.
9. A similar position is adopted by other approaches that acknowledge the importance of investigating the interplay of lexis and grammar (colligation and related notions) in linguistic analysis (e.g., Pattern Grammar or Construction Grammar).

References

Biber, D. (1988), *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (1993), "The multi-dimensional approach to linguistic analyses of genre variation: an overview of methodology and findings," *Computers and the humanities*, 26: 331–345.

Biber, D. (1995), *Dimensions of register variation: a cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biber, D., S. Johansson, and G. Leech (1999), *Longman grammar of spoken and written English*. Harlow: Longman.

Chakrabarti, S. (2003), *Mining the Web: discovering knowledge from hypertext data*. Boston: Morgan Kaufmann.

Gries, St. Th. (2006), "Exploring variability within and between corpora: some methodological considerations," *Corpora,* 1(2): 109–151.

Halliday, M. A. K. (1985), *Spoken and written language*. Victoria: Deakin University Press.

Halliday, M. A. K. (2004), *Introduction to functional grammar*. London: Edward Arnold.

Halliday, M. A. K. (2005 [1991]), "Towards probabilistic interpretations," in: J. J. Webster (ed.) *Computational and quantitative studies.* London: Continuum, 42–62.

Halliday, M. A. K., and R. Hasan (1989), *Language, context and text: aspects of language in a social-semiotic perspective*. Oxford: Oxford University Press.

Halliday, M. A. K., and J. R. Martin (1993), *Writing science, literacy and discursive power.* London: The Falmer Press.

Halliday, M. A. K., A. McIntosh, and P. Strevens (1964), *The linguistic sciences and language teaching*. London: Longman.

Holtz, M., and E. Teich (in preparation), "Scientific registers in contact: an exploration of the lexico-grammatical properties of interdisciplinary discourses."

Kilgarriff, A. (2001), "Comparing corpora," *International journal of corpus linguistics*, 6(1): 1–37.

Mair, C. (2006), *Twentieth-century English: history, variation, and standardization*. Cambridge: Cambridge University Press.

Mair, C. (2009), "Corpora and the study of recent change," in: A. Lüdeling and M. Kytö (eds.) *Corpus linguistics*. Berlin: Mouton de Gruyter: 1109–1125.

Manning, C., and H. Schütze (1999), *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Matthiessen, C. M. I. M. (1993), "Register in the round, or diversity in a unified theory of register," in: M. Ghadessy (ed.) *Register analysis: theory and practice.* London: Pinter, 221–292.

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik (1985), *A comprehensive grammar of the English language*. London: Longman.

Schmid, H. (1994), "Probabilistic part-of-speech tagging using decision trees," in: *Proceedings of the 1st international conference on new methods in language processing*, 44–49.

Schwarz, L., S. Bartsch, R. Eckart, and E. Teich (2008), "Exploring automatic theme identification: a rule-based approach," in: A. Storrer, A. Geyken, A. Siebert, and K.-M. Würzner (eds.) *Text resources and lexical knowledge: selected papers from the 9th conference on natural language processing.* Berlin/New York: Mouton de Gruyter, 15–26.

Sebastiani, R. (2002), "Machine learning in automated text categorization," *ACM computing surveys*, 34(1): 1–47.

Teich, E. (2003), *Cross-linguistic variation in system and text: a methodology for the investigation of translations and comparable texts*. Berlin/New York: Mouton de Gruyter.

Teich, E. (2009), "Linguistic computing," in: M. A. K. Halliday and J. A. Webster (eds.) *Companion to systemic functional linguistics*. London: Equinox, 113–127.

Witten, I. H., and F. Eibe (2005), *Data mining: practical machine learning tools and techniques*. Boston: Morgan Kaufmann.

Automated learning of appraisal extraction patterns

Kenneth Bloom and Shlomo Argamon

Linguistic Cognition Lab
Dept. of Computer Science
Illinois Institute of Technology

Abstract

This paper describes a grammatically motivated system for extracting opinionated text. A technique for extracting appraisal expressions has been described in previous work, using manually constructed syntactic linkages to locate targets of the opinions. The system extracts attitudes using a general lexicon—and some candidate targets using a domain specific lexicon—and finds additional targets using the syntactic linkages. In this paper, we discuss a technique for automatically learning the syntactic linkages from a list of all extracted attitudes and the list of candidate targets. The accuracy of the new learned linkages is comparable to the accuracy of the old manual linkages.

1. Introduction

Many traditional data mining tasks in natural language processing focus on extracting and mining topical data. In recent years, the natural language community has recognized the value in analyzing opinions and emotions expressed in free text, developing the field of sentiment analysis to research applications of opinionated text and methods for extracting it. While early applications focused on review classification (Pang and Lee, 2004) or sentence classification (Seki et al., 2007), many recent applications involve opinion mining in ways that require a structured view of the opinions expressed in a text—for example Archak, Ghose, and Ipeirotis' (2007) application of sentiment analysis to analyzing product pricing.

We have proposed (Bloom, Garg, and Argamon, 2007) that a fundamental task in sentiment analysis is the extraction of *appraisal expressions*, the basic grammatical structure expressing a single opinion. In their most basic form, appraisal expressions consist of three common parts, including an *attitude* which states the nature of the opinion, a *target* which the opinion is about, and a *source* who expresses the opinion. Other parts may also be present, such as a second target, when the attitude is a comparative adjective. An appraisal expression is not necessarily contained in a single sentence, and parts of an expression may be elided or filled by anaphoric references to other sentences.

Our work on appraisal expression extraction began with shallow parsing attitudes and applying modifiers to compositionally represent the meaning of adjectival attitude phrases (Whitelaw, Garg, and Argamon, 2005). We later developed a technique for finding targets of opinions using shallow parsing and

dependency parse paths (Bloom, Garg, and Argamon, 2007). The technique discussed in that paper used a lexicon of common opinion targets for a given domain, and a hand-built list of dependency parse paths to link these potential targets to attitudes and to identify new targets not included in the lexicon. In this paper, we present a technique for learning these paths from the extracted attitudes and the partial lexicon of opinion targets. We demonstrate that the effectiveness of this learning technique is comparable to the manually constructed list of attitudes used in the previous paper.

2. What is appraisal?

Appraisal theory (Martin and White, 2005), based in Systemic-Functional Linguistics, deals with three grammatical systems that convey opinion. The ENGAGEMENT system deals with how writers position statements with respect to other possible statements on the same subject (such as admitting or discounting the possibility of other views). The ATTITUDE system deals with direct evaluations of people, objects, and facts. The GRADUATION system deals with the way evaluations conveyed by the ATTITUDE system are affected by modifiers. We model one piece of the GRADUATION system—the attribute of *force*, or the strength of an attitude.

In broad outline, the ATTITUDE system classifies the grammatical type of the opinion into one of three types: *affect* refers to an emotional state (e.g., 'happy' or 'angry'), and is the most explicit type of appraisal, being centered on the person experiencing the emotion. *Appreciation* evaluates the intrinsic qualities of an object (e.g., 'it's beautiful'), while *judgment* evaluates a person's behavior in a social context (e.g., 'he's evil'). The ATTITUDE system also deals with the orientation of opinions, determining whether they are positive or negative.

The orientation of an attitude, which is whether the attitude is negative or positive, is also an obvious part of appraisal theory, and determining the orientation of an attitude is the essential task in sentiment analysis. Our software system is designed to extract and analyze *attitude groups*, which are realizations of the ATTITUDE system, the GRADUATION system, and orientation.

Systemic-functional systems specify a network of choices that writers can make about the meanings they wish to convey in their writing, and these choices impose constraints on the text. The ATTITUDE system imposes constraints on the lexis used to express opinions, properties of the target, and the grammatical locations of other parts of the appraisal expression. While Martin and White do not discuss grammatical locations at all, Hunston and Sinclair (2000) explored the structural patterns by which adjectival evaluation is expressed in text, but with no relation to appraisal theory. Bednarek (2006) has done some work in connecting these two theories, explaining how differences in the attitude expressed affect the availability of different structural patterns to convey the attitude.

3. Related work

The technique presented for extracting appraisal expressions presented here is based on our previous work (Bloom, Garg, and Argamon, 2007). We first identified opinions from a lexicon, and then used a list of 29 linkage specifications to find targets, based on the position of the attitudes. Work by Popescu and Etzioni (2005) takes a similar approach, albeit with some differences. They first find explicit features of the product under review, by using simple extraction patterns with their KnowItAll information extraction system. Afterward, they use a short list of ten grammatical linkages to find opinion words, and they use relaxation labeling to assess the orientation and force of those opinion words. It appears that the learning technique presented in this paper should be useful with both extraction systems.

Prior work that relates to learning linkage specifications comes from information extraction, where several techniques have been proposed for learning extraction patterns for binary relations. Because the nature of information extraction generally deals with relations that occur much less frequently than the attitude-target relation, many of these techniques are concerned with learning much longer lists of high-precision, low-recall patterns. These patterns are frequently based on more specific features than purely syntactic linkages, such as the text surrounding the entities to be extracted.

Such is the case in DIPRE (Brin, 1998) and its successor Snowball (Agichtein and Gravano, 2000), which start with small lists of around ten examples of the target relation and learn patterns of the form <*text-before, slot-1, text-between, slot-2, text-after*>. They iterate several times, generating new lists of patterns and using those to generate new lists of high-confidence seeds. The two methods differ mainly in how they compute confidence and in how they represent the text in the patterns.

There are several supervised approaches to binary relation learning. The closest to our approach is that of Miller et al. (2000), who augment the phrase types in a phrase-structure syntactic parser to recognize grammatical phrases as corresponding to particular relations and slots.

Kambhatla (2004) has developed a method for predicting the type of relationship between pairs of mentions using a multi-class classifier. This technique extracts all entities using a named entity recognizer, then trains a multi-class maximum entropy classifier to predict one of 24 relation types, or assign a "no relation" option. Features used in the classification include words in and between the mentions, entity types of the mentions, dependency parse features, and phrase-structure parse features.

4. Appraisal extraction system

Our attitude extraction system operates in three stages. In the first stage, it identifies attitudes and some candidate targets by shallow parsing. In the second

stage, it links attitudes to candidate targets and identifies new targets by syntactic position. In the third stage, it disambiguates between multiple possible appraisal expressions that can be constructed from a single attitude group.

4.1 Identifying attitudes and targets

The first phase is identifying attitudes and candidate targets by shallow parsing, a process we call *chunking*. In shallow parsing, the computer works to identify the beginnings and ends of a certain type of phrase (like a noun phrase or verb phrase) without the use of a full phrase structure parse. The goal of the chunker is to shallow parse in order to find phrases of a certain type, and compute attributes of the phrases it finds, modeling the behavior of modifiers within the phrase.

When extracting attitudes, the chunker begins with a lexicon of nominal and adjectival appraisal head words, which define initial values for attributes of the attitudes these convey. These attributes include the attitude type and the orientation and force of the attitude. These lexicon entries can be ambiguous, specifying multiple possibilities for type of attitude conveyed by the headword. Other attributes of the attitude can also vary with the attitude type; for example, the word *devious* can be a realization of negative propriety or positive capacity, depending on context.

The chunker looks for occurrences of these in the text, and upon finding them it looks leftward (taking advantage of English word order) to attach modifiers to the words and update the values of the attributes according to the modifiers. This process is described by Whitelaw, Garg, and Argamon (2005).

The chunker is also used to identify potential targets using a domain specific target lexicon and type taxonomy. The lexicon contains many of the most common words used as targets in a specific domain. For example, when working with reviews of digital cameras, the lexicon contains a list of the parts of a camera. When working with movie reviews, it contains general terms referring to aspects of movie-making and marketing. It also contains a list of actor, director, and character names from the movie that the particular review discusses, all slurped from IMDb. The chunker finds all of the target groups matching words specified in the lexicon, but the target lexicon does not have enough coverage to find all of the target groups that are actually present in the text. The associator (described in the next section) is designed to find additional target groups based on their syntactic position.

We perform chunking with several other lexicons as well, to give us additional information about the target. For example, we have a lexicon that captures the DEICTICITY system of noun phrases (see Halliday and Matthiessen, 2004). A named entity recognizer may also be run at this stage if it is appropriate for the corpus. The associator (described in the next section) will gather the various kinds of chunks associated with a given target.

4.2 Linking attitudes to targets

After finding attitude groups and candidate targets, the system links attitudes to targets. There are two goals for this phase: (1) finding a nominal group that is the textual representation of the target, and (2) computing the values of attributes describing that target.

The associator finds the target phrase by following paths through a dependency parse of the sentence containing an attitude. In a preprocessing step, our corpora are parsed using the Stanford dependency parser (de Marneffe, MacCartney, and Manning, 2006). The associator contains a ranked list of paths through the dependency parse; these paths specify that certain kinds of links must be traversed in a certain order. These paths are called *linkage specifications*. For each attitude in the corpus, the system looks for *linkages*, paths through the dependency tree that connect any word in the attitude to what will become the last word of the target phrase and that match one of the linkage specifications. Upon finding a word in a suitable syntactic position, the associator performs shallow parsing, looking to the left to build a noun phrase that ends in the located word.

To compute the values of attributes describing the target, the associator gathers various kinds of target chunks, such as the candidate target, the DEICTICITY system, and named entities mentioned in the previous section. When any of these chunks overlaps with the nominal group found by the associator, it will be considered as a part of the target, and the values of its attributes will be used as attributes of the target.

For example, take the linkage specification

$$target \xrightarrow{nsubj} x \xleftarrow{dobj} y \xleftarrow{amod} attitude$$

When we apply it to the sentence *The Matrix was a good movie*, the chunker finds *good* as an attitude. The associator then finds the word *Matrix* as shown in Figure 1. Through shallow parsing, we determine that this is part of the phrase *The Matrix*, which is the target of the evaluation. Since the phrase *The Matrix* is in the target lexicon because it is the name of the movie, its attributes are copied and used as attributes of the target.

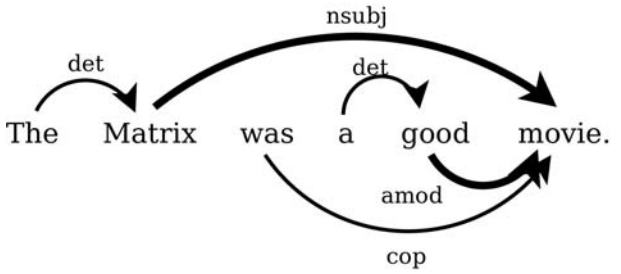


Figure 1. An example of how the associator matches a target.

4.3 Disambiguation

At this point, the appraisal expression for a given attitude group may have multiple interpretations found by the computer, with different linkages used, as well as ambiguities in the attitude type or other attributes of the extracted appraisal.

The resolution of ambiguities created by different matching linkage specifications is resolved by a postprocessing step built into the associator. Any linkage which connects to a candidate target (from the domain-specific lexicon chunked previously) is given higher priority than linkages that do not include a candidate target. Where multiple linkages all connect to candidate targets, or where there are multiple linkages but none connect to candidate targets, one that comes earlier in the list of linkage specifications has higher priority than one that comes later. The postprocessing step retains only the highest priority linkage for each attitude group.

Where an attitude group had an ambiguous interpretation (for example, the word *devious* described above), a probabilistic Expectation-Maximization learner is used to learn the probability of different attitude types, given the attributes of the target and the linkage specification used to connect the attitude to the target. Since most attitude groups are not double-coded for attitude type, the system bootstraps from the singletons and learns the parameters required to disambiguate the ambiguous instances.[2]

5. Learning linkages

To make a system which encodes a lot of knowledge feasible for consumer use, it is important to develop methods whereby the various kinds of knowledge used in the system can be learned quickly, without requiring many man-hours of work to adapt the system to new domains. To that end, we have developed a technique for learning linkage specifications from a dependency-parsed but untagged corpus and a seed lexicon of targets and attitudes.

To operate the learner, we begin by running the chunker to find all attitude groups using the attitude lexicon, and to find a collection of potential target groups using the domain-specific target lexicon. As mentioned above, the chunker finds all of the target groups matching words specified in the lexicon, but the target lexicon does not have enough coverage to find all of the target groups that are actually present in the text.

The linkage learner looks at each sentence and considers all possible attitude-target pairs in that sentence. It computes the path through the dependency parse tree which connects the last word of the attitude to the last word of the target, and adds it to the list of linkage specifications found by the learner. For each linkage specification, a count B is maintained, counting the number of times a specification was seen connecting both a chunked attitude and a chunked target. Linkage specifications with more than five links in them are not considered by the learner.

After finding all possible linkage specifications, the learner tests the specifications in descending order of *B* to compute more statistics for the linkage specifications. (Because this is a time consuming process, we asked the learner to compute only full statistics for the top 100 linkage specifications, but this number is configurable.) It computes *T'*, the number of times the linkage specification was seen with an attitude but no candidate target; *A'*, the number of times it was seen connecting a candidate target, but no attitude; and *N*, the number of times it was seen with neither an attitude nor a candidate target. These four statistics can be used by scoring metrics to find the best linkage specifications to be used by the associator. Though we don't use all four, there is no extra cost to computing all four for use in further research.

Finally, to select appropriate specifications for use in the associator, we apply a scoring metric to pick the top rules. We consider the value

$$\frac{B}{B+T'}$$

to be the *confidence* of the linkage specification, which measures its tendency to connect to candidate targets in places where it's known to connect to an attitude. (This is similar to the situation when the associator is actually run.) We cannot use the confidence as our scoring metric, however, as it tends to favor longer, more specific linkage specifications. As a result, many appraisal expressions are extracted where no target is found because no syntactic path from the attitude group matches any of the linkage specifications.

To adjust for this, we use the scoring metric

$$B \cdot \frac{B}{B+T'}$$

to give high weighting to linkage specifications that are used frequently, but lower their score when they have lower confidence. This metric tends to slightly favor shorter, more general linkage specifications, while still giving high confidence linkage specifications a boost.

6. Corpora and lexicons

The lexicon used to identify attitudes is based on the lexicon originally developed for our previous work (Bloom, Garg, and Argamon 2007). Since the publication of that paper, the authors have extended it to include nominal appraisal head words, and modifiers which modify appraisal in nominal groups. The current version of the lexicon includes 207 modifiers, and 3,814 appraisal head words (2,108 are adjectival appraisal, and 1,706 are nominal appraisal). The lexicon is hand-built, with words culled from several sources, including sample movie reviews, Martin and White's (2005) samples, WordNet synset expansion (Miller, 1995), and the lexicons used in the General Inquirer (Stone et al., 1966).

We tested the learner on two corpora of reviews, each with its own domain-specific target lexicon.

The first corpus is a collection of user product reviews taken from epinions.com and supplied in 2004 for research purposes by Amir Ashkenazi of shopping.com. The base collection contains reviews for three types of products: baby strollers, digital cameras, and printers. Each review has a numerical rating (1–5); based on this, we labeled positive and negative reviews in the same way as Pang and Lee (2004) did for the movie reviews corpus. The products corpus comprises 15,162 documents (11,769 positive documents, 1,420 neutral documents, and 1,973 negative documents), averaging 442 words. There are 905 reviews of strollers, 5,778 reviews of ink-jet printers, and 8,479 reviews of digital cameras, covering 516 individual products.

Generic target lexicons were constructed by starting with a small sample of the kind of reviews that the lexicon would apply to. We examined these manually to find generic words referring to appraised things to serve as seed terms for the lexicon and used WordNet to suggest additional terms to add to the lexicon.

The second corpus is the standard publicly available collection of movie reviews constructed by Pang andand Lee (2004) for the review classification task. This standard testbed consists of 1,000 positive and 1,000 negative reviews, all taken from the IMDB movie review archives.[2] Reviews with neutral scores (three stars out of five) were removed by Pang and Lee, giving a data set with only clearly positive and negative reviews. The average document length in this corpus is 764 words, and 1,107 different movies are reviewed.

For the IMDB corpus, we constructed a target lexicon of generic terms related to aspects of movie-making and marketing. Since movie reviews often refer to the specific contents of the movie under review by proper names (of actors, the director, etc.), we also automatically constructed an additional lexicon for each movie in the corpus, based on lists of actors, characters, writers, directors, and companies listed for the film at imdb.com. This additional lexicon differs from the output of a named entity recognizer (NER) because it breaks down the extracted names into fine levels of detail denoting their particular role in the movie. (Most NER packages stop at the level of person, organization, or location.) Each movie-specific lexicon was used only for processing reviews of the movie it was generated for, so the system had no specific knowledge of terms related to other movies during processing.

From the experience that we have gained with the corpora while evaluating them, we have determined that the product reviews contain simpler appraisals, less creative metaphor, less varied syntactic structure (specifically a noticeably higher incidence of situations where the attitude is an adjectival modifier of its target), and less varied attitude types (a strong emphasis on *quality*, a subtype of *appreciation*). The movie reviews corpus contains more metaphor and more variation in syntactic structure and attitude type. A particular hallmark of evaluation in movie reviews is the presence of evaluation written into

in the plot of a movie by the movie's makers and summarized by the reviewer as part of a plot summary (for example, references to evil characters).

7. Results

We evaluated the linkage learner by comparing the performance of its learned linkage specifications against a manually constructed list of linkage specifications based on those used in our previous work (Bloom, Garg, and Argamon, 2007). Beginning with their list of linkage specifications, we added 12 new linkage specifications specific to nominal appraisal at varying priorities in the list.

We operated the linkage learner on each of the corpora separately, and we evaluated each set of learned linkage specifications on the corpus it had been learned from.

One of the authors rated the extractions. He rated 150 appraisal expressions in each of four experiments—a manual linkage experiment and a learned linkage experiment for each of two corpora. He was presented with extracted appraisal expressions, with the target and attitude bracketed in the sentence from which they were drawn. He evaluated them for several criteria:

- *Appraisal*. Indicates whether the rater thought the extracted attitude group actually conveyed an attitude in context.
- *HumTgt*. Indicates whether the rater could identify a target of the appraisal in the presented sentence. If the target was anaphorically referenced from another sentence or was elided in the presented sentence, the target was not in the presented sentence.
- *Correct*. If the rater was able to identify a target in the presented sentence, did the computer find the correct target?
- *Percent*. The percentage of appraisal expressions for which he could identify the target where the computer also got them correct.

The results are presented in Table 1. The linkage specifications learned by our learner tended to perform comparably to the manually constructed linkages previously used in this system. Both systems tended to suffer from the same kinds of errors. These errors include spurious appraisal, parser errors, selection of the wrong linkage where several linkages matched, and appraisals that don't have targets.

This last situation occurs particularly with nominal appraisal, where an appraisal may be an anaphoric reference to its own target; for example, in the sentence *This **[problem]** outweighs all of the positive points of this product, as far as I'm concerned*, the attitude *problem* is a reference to its own target, which is described more fully in a previous sentence. Another similar situation is where a quality is mentioned in the abstract and not associated with any target; for example, *but the laughs are built on **[discomfort]** and **[embarrassment]**, not on any intrinsic humor from within the story itself.*

Corpus	Experiment	Appraisal	HumTgt	Correct	Percent
Products	manual	117	105	73	69%
	learned	116	105	68	64%
Movies	manual	128	101	63	62%
	learned	116	89	50	56%

Table 1. Results from accuracy evaluation of extraction with learned and manually constructed linkages

8. Conclusion

We have described a system for extracting opinion targets that uses grammatical linkages to find suitable phrases, and heuristic postprocessing to select the correct target from several candidates. We have presented a method for learning the grammatical linkages necessary to find opinion targets starting from a seed lexicon. A manual evaluation shows that automatically learned linkages perform comparably to a manually constructed list of linkages used for the same purpose.

Immediate future work includes developing a more flexible machine-learning disambiguator that can select the correct link using local context and handle the other types of disambiguities about attitude type that we have discussed. Ideally, we would like to develop a new disambiguator that considers multiple interpretations of an appraisal expression (representing ambiguity in all of these attributes), and learns to rank them and select the highest ranked interpretation. This would eliminate the need for a postprocessing step in the associator, and would inherit the functionality of the probabilistic disambiguator.

9. Notes

1. A third type of ambiguity that is not yet resolved is where a particular candidate target had ambiguity among its domain-specific types. For example, when a movie is written and directed by the same person, we wish to disambiguate whether an appraisal of that person is an appraisal of him in his capacity as the writer or director of the movie.
2. See <http://www.cs.cornell.edu/people/pabo/movie-review-data>.

References

Agichtein, E., and L. Gravano (2000), "Snowball: extracting relations from large plain-text collections," in: *Proceedings of the 5th ACM international conference on digital libraries*. ACM. Online at: <http://www1.cs.columbia.edu/~gravano/Papers/2000/dl00.pdf>.

Archak, N., A. Ghose, and P. G. Ipeirotis (2007), "Show me the money: deriving the pricing power of product features by mining consumer reviews," in: P.

Berkhin, R. Caruana, and X. Wu (eds.) *KDD '07: Proceedings of the 13th ACM international conference on knowledge discovery and data mining*. ACM, 56-65. Online at: <http://pages.stern.nyu.edu/~aghose/kdd2007.pdf>.

Bednarek, M. (2006), "Language patterns and ATTITUDE," unpublished m.s.

Bloom, K., N. Garg, and S. Argamon (2007), "Extracting appraisal expressions," in: S. L. Sidner, T. Schultz, M. Stone, and C. Zhai (eds.) *Proceedings of the human language technology conference of the North American chapter of the ACL*. ACL, 308–315. Online at: <http://www.aclweb.org/anthology/N/N07/N07-1039.pdf>.

Brin, S. (1998), "Extracting patterns and relations from the World Wide Web," in: P. Atzeni, A. O. Mendelzon and G. Mecca (eds.) *The World Wide Web and databases*. New York: Springer, 172–183.

De Marneffe, M.-C., B. MacCartney, and C. D. Manning (2006), "Generating typed dependency parses from phrase structure trees," in: *Proceedings of the 5th international conference on language resources and evaluation*. Online at: <http://www.sdjt.si/bib/lrec06/>.

Halliday, M. A. K., and C. M. I. M. Matthiessen (2004), *An introduction to functional grammar*. London: Edward Arnold.

Hunston, S., and J. Sinclair (2000), "A local grammar of evaluation," in: S. Hunston and G. Thompson (eds.) *Evaluation in text: authorial stance and the construction of discourse*. Oxford, UK: Oxford University Press, 74–101.

Kambhatla, N. (2004), "Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations," in: *Proceedings of the ACL 2004 on interactive poster and demonstration sessions*. ACL, 178–181.

Martin, J. R., and P. R. R. White (2005), *The language of evaluation: appraisal in English*. London: Palgrave.

Miller, S., H. Fox, L. A. Ramshaw, and R. M. Weischedel (2000), "A novel use of statistical parsing to extract information from text," in: *Proceedings of the 6th applied natural language processing conference*. ACL, 226–233. Online at: <http://acl.ldc.upenn.edu/A/A00/A00-2030.pdf>.

Miller, G. A. (1995), "WordNet: a lexical database for English," *Communications of the ACM*, 38(11): 39–41.

Pang, B., and L. Lee (2004), "A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts," in: *Proceedings of, 42nd meeting of the ACL*. ACL, 271-278. Online at: <http://www.cs.cornell.edu/people/pabo/papers/acl04_cutsent.pdf>.

Popescu, A.-M., and O. Etzioni (2005), "Extracting product features and opinions from reviews," in: *Proceedings of the human language technology conference/conference on empirical methods in natural language processing*. ACL, 339-346. Online at: <http://turing.cs.washington.edu/papers/emnlp05_opine.pdf>.

Seki, Y., D. K. Evans, L.-W. Ku, H.-H. Chen, N. Kando, and C.-Y. Lin (2007), "Overview of opinion analysis pilot task at NTCIR-6," in: *Proceedings of the 6th NTCIR workshop*, 265-278. Online at: <http://nlg3.csie.ntu.edu.tw/conference_papers/ntcir6opinion.pdf>.

Stone, P. J., D. C. Dunphy, M. S. Smith, and D. M. Ogilvie (1966), *The general inquirer: a computer approach to content analysis*. Cambridge: MIT Press.

Whitelaw, C., N. Garg, and S. Argamon (2005), "Using appraisal taxonomies for sentiment analysis," in: O. Herzog, H.-J. Scheck, N. Fuhr, A. Chowdhury, and W. Teiken (eds.) *Proceedings of the 14th ACM international conference on information and knowledge management*. ACM, 625–631. Online at: <http://lingcog.iit.edu/doc/appraisal_sentiment.pdf>.