Spoken and written register variation in Spanish: A multi-dimensional analysis[1]

Douglas Biber, Mark Davies,
James K. Jones, and Nicole Tracy-Ventura[2]

Abstract

There have been few comprehensive analyses of register variation conducted in a European language other than English. Spanish provides an ideal test case for such a study: Spanish is a major international language with a long social history of literacy, and it is a Romance language, with interesting linguistic similarities to, and differences from, English. The present study uses Multi-Dimensional (MD) analysis to investigate the distribution of a large set of linguistic features in a wide range of spoken and written registers: 146 linguistic features in a twenty-million words corpus taken from nineteen spoken and written registers. Six primary dimensions of variation are identified and interpreted in linguistic and functional terms. Some of these dimensions are specialised, without obvious counterparts in the MD analyses of other languages (e.g., a dimension related to discourse with a counterfactual focus). However, other Spanish dimensions correspond closely to dimensions identified for other languages, reflecting functional considerations such as interactiveness, personal stance, informational density, argumentation, and a narrative focus.

1. Introduction

The importance of register as an explanatory factor for linguistic variation has been increasingly recognised over the past two decades. Numerous studies in functional linguistics, which focus on the interaction of discourse and grammar, have documented how spoken and written register differences help to explain the patterns of variation for a linguistic feature (e.g., Prince, 1978; Tottie, 1991; Collins, 1991; Sigley, 1997; Oh, 2000; Kaltenböck, 2005). Most of these studies use corpus-based analysis to show how characteristics of the textual context interact with register

[1] Work on this project was supported by National Science Foundation Research Grant #BCS-0214438.

[2] *Correspondence to*: Douglas Biber, *e-mail*: douglas.biber@nau.edu
address: Department of English, Liberal Arts Building, Room 140, P.O. Box 6032, Northern Arizona University, Flagstaff, Arizona 86011-6032, USA

differences, so that strong patterns of use in one register often represent only weak patterns in other registers. More recently, the *Longman Grammar of Spoken and Written English* (Biber *et al.*, 1999) provides a comprehensive grammatical description of English that documents how most grammatical features and variants are distributed in systematic ways across spoken and written registers.

In fact, several scholars have argued that register variation is inherent in human language: a single speaker will make systematic choices in pronunciation, morphology, word choice, and grammar reflecting a range of situational factors. For example:

> 'each language community has its own system of registers ... corresponding to the range of activities in which its members normally engage'
>
> (Ure, 1982: 5)

> 'register variation, in which language structure varies in accordance with the occasions of use, is all-pervasive in human language'
>
> (Ferguson, 1983: 154)

> 'no human being talks the same way all the time... At the very least, a variety of registers and styles is used and encountered.'
>
> (Hymes, 1984: 44)

Surprisingly, despite the demonstrated importance of register variation, there have been few comprehensive analyses of the register differences in a language. This gap is due mostly to methodological difficulties: until recently, it has been unfeasible to analyse the full range of texts, registers, and linguistic characteristics required for comprehensive analyses of register variation. However, with the availability of large on-line text corpora and computational analytical tools, such analyses have become possible. Multi-Dimensional (MD) analysis – the research methodology applied in the present study – is a corpus-based approach developed for the comprehensive analysis of register variation.

More specifically, MD analyses describe the basic patterns of linguistic variation among spoken and written registers, and the ways (and extent) in which any two registers are similar or different linguistically (see, for example, Biber, 1988, 1995; Conrad and Biber, 2002). This research approach is based on computational analysis of a large text corpus to identify the most important patterns of linguistic co-occurrence: the "dimensions". Each dimension comprises a distinct set of co-occurring linguistic features, and each has distinct functional underpinnings. Registers can be compared in this multi-dimensional space, enabling empirical analysis of both the extent and the ways in which any two registers are different.

There have been numerous MD analyses of English, from both synchronic and diachronic perspectives, considering a wide range of

general as well as more specialised registers. There have also been major analyses of Korean and Somali, and more restricted analyses of other languages. These data-driven analyses have resulted in many unanticipated findings, including: major differences among "oral" and "literate" registers but no absolute differences between speech and writing (Biber, 1988); a fundamental distinction between the linguistic complexity profiles of spoken registers (which differ in extent but not kind) and written registers (which exploit the full spectrum of linguistic variation; Biber, 1992); surprising similarities in the underlying multi-dimensional structure of English, Korean, and Somali, complemented by specific differences reflecting the communicative priorities of each culture (Biber, 1995: chapter 7); and dramatic historical shifts in the patterns of register variation in English and Somali (Biber, 1995: chapter 8).

Surprisingly, there have been fewer MD analyses of register variation in other European languages. In the present study, we help to fill this gap by undertaking an MD analysis of register variation in Spanish. Spanish is an ideal complement to previous analyses. From a cross-linguistic and cross-cultural perspective, Spanish has many interesting points of similarity to, and difference from, English, Korean, and Somali. Spanish is a major international language used for a full range of interpersonal as well as institutional functions. It has a long social history of literacy which has been influenced by the role of an official language academy. It is a Romance language, with interesting linguistic similarities to, and differences from, English. And there are important differences in the preferred conversational styles and rhetorical priorities in writing. Given these characteristics, we had every reason to expect that a MD analysis of register variation in Spanish would produce findings that were as surprising and interesting as earlier MD analyses. The following sections show that these expectations were borne out by the actual research findings.

2. Background: previous corpus-based studies of Spanish

Although there have been few comprehensive analyses of spoken and written register variation in Spanish, there have been numerous corpus-based studies. Until the late 1990s, there were two corpora that were the focus of most variation studies of Spanish. The first was the *Habla Culta*, comprising 2.5 million words of transcribed interviews and conversations from eleven cities in Latin America and Spain (Lope Blanch, 1977, 1991; this corpus is currently available as part of the *Corpus del Español*; see below). The second consists of three, one-million words sub-corpora created at the Universidad Autónoma de Madrid: the *Corpus oral de referencia de la lengua española contemporánea*, the *Corpus lingüístico de referencia de la lengua española en Argentina*, and the *Corpus lingüístico de referencia de la lengua española en Chile* (Ballester and Santamaría,

1993; Marcos-Marín, 1996). Since the late 1990s, several other large public domain corpora have become available, including the *Biblioteca Virtual* (http://www.cervantesvirtual.com); *CORDE* (historical Spanish) and *CREA* (modern Spanish), both from the Real Academia Española (http://www.rae.es); and the 100 million words *Corpus del Español*, the first large, tagged corpus of Spanish (http://www.corpusdelespanol.org).

Many studies of syntactic variation in Spanish have focused on dialect differences, mostly using the *Habla Culta* (see, e.g., Lope Blanch, 1977, 1991; Davies, 1995, 1997, 2003; De Mello, 1992a, 1992b, 2002, 2004; Butler, 1992; Sedano, 1994b; Ocampo, 1995). Other studies have taken a register perspective. Several of these have described salient linguistic characteristics of a single register, including academic texts (Gibbons, 1999), newspapers (Thibault, 1987), parliamentary texts (Alcaide Lara, 1999), and conversation (De Mello, 1995). In addition, several studies have compared the linguistic characteristics of spoken and written registers in Spanish: Butler (1998) on collocations; Arce Castillo (1999) on intensifiers; Davies (1995, 1997) on causatives and clitic climbing; Brizuela, Andersen, and Stallings (1999) on discourse markers.

There have been two previous multi-dimensional studies of register variation in Spanish. In the first, Sáiz (1999) built parallel corpora of English and Spanish texts (for example, the Xerox ScanWorX User's Guide, translated into both languages), and then undertook independent MD analyses of both sub-corpora. The study focused primarily on part-of-speech and simple grammatical distinctions (e.g., plural nouns, present tense verbs), resulting in five dimensions being identified for both languages. These dimensions were for the most part similar in their underlying functions across the two languages, and the parallel registers were also similar in many respects.

Parodi (2005) presents a more developed MD analysis of Spanish, based on the distribution of sixty-five linguistic features in a 1.5 million words corpus from three major registers taken from technical-professional high schools in Valparaíso, Chile: technical/scientific texts, fictional literature, and oral interviews. This study uncovered five major dimensions, interpreted as "contextual and interactive focus", "narrative focus", "commitment focus", "modalizing focus", and "informational focus". The first two of these are very similar to important dimensions in the MD analysis of English (Biber, 1988), while the others are more specialised for these registers in Spanish.

The present study complements previous studies of linguistic variation in Spanish, including the three previous MD studies, by investigating the distribution of a large set of linguistic features in a wide range of spoken and written registers: 146 linguistic features in a twenty-million words corpus taken from nineteen spoken and written registers. As the following sections show, the analysis identifies five major linguistic dimensions, with each comprising a distinct set of co-occurring linguistic features which reflect different underlying communicative functions. Some

of these are similar to the dimensions of variation identified in the analyses of English and other languages, supporting the possibility of universal dimensions of register variation. However, other dimensions are unique to the present analysis, reflecting the particular linguistic and communicative resources of Spanish.

3. Methodology

3.1 Conceptual introduction to the multi-dimensional approach

Multi-Dimensional (MD) analysis was developed as a corpus-based methodological approach to, (i) identify the salient linguistic co-occurrence patterns in a language, in empirical/quantitative terms, and (ii) compare registers in the linguistic space defined by those co-occurrence patterns. The approach was first used by Biber (1985, 1986) and then developed more fully in Biber (1988).

The co-occurrence patterns comprising each dimension are identified quantitatively. That is, a factor analysis is used to identify the sets of linguistic features that frequently co-occur, based on the distributions of linguistic features in a large corpus of texts. Qualitative analysis is also required to interpret the functions associated with each set of co-occurring linguistic features. Thus, the dimensions of variation have both linguistic and functional content. The linguistic content of a dimension comprises a group of linguistic features (e.g., nominalisations, prepositional phrases, attributive adjectives) that co-occur with a high frequency in texts. Based on the assumption that co-occurrence reflects shared function, these co-occurrence patterns are interpreted in terms of the situational, social, and cognitive functions most widely shared by the linguistic features. That is, linguistic features co-occur in texts because they reflect shared functions.

3.1.2 Overview of analytical steps

Following the methodology used in the MD analyses of other languages (see, e.g., Biber, 1995: chapter 5; Conrad and Biber, 2001: chapter 2), the analysis here required eight main methodological steps:

i) An appropriate corpus was designed and constructed, based on the major research questions and goals of the project,

ii) Research was conducted to identify the linguistic features to be included in the analysis, together with functional associations of the linguistic features,

iii) Computer programs were developed for automated grammatical analysis, to identify – or "tag" – all relevant linguistic features in texts. Extensive testing and revision of the tagger resulted in a set of features that were identified with a high degree of accuracy,

iv) The entire corpus of texts was tagged automatically by computer,

v) Additional computer programs were developed and run to compute normed frequency counts of each linguistic feature in each text of the corpus,

vi) The co-occurrence patterns of linguistic features were identified through a factor analysis of the frequency counts,

vii) The factors from the factor analysis were interpreted functionally as underlying dimensions of variation, and

viii) Dimension scores for each text were computed, and the mean dimension scores for each register were then compared to analyse the salient linguistic similarities and differences between registers.

3.2 The corpus used for the analysis

The corpus used for the study comes from the twentieth-century component of the NEH-funded Corpus del Español (www.corpusdelespanol.org; see Davies, 2002). The Corpus del Español incorporates texts from many other existing Spanish corpora, including Habla Culta (Lope Blanch 1977, 1991), the Corpus oral de referencia de la lengua española contemporánea, the Corpus lingüístico de referencia de la lengua española en Argentina, the Corpus lingüístico de referencia de la lengua española en Chile (Ballester and Santamaría, 1993; Marcos-Marín, 1996), and *The Biblioteca Virtual* (http://www.cervantesvirtual.com). In addition, we added a sample of forty academic research articles in science and the humanities, downloaded from on-line sources.

For the present study, we categorised all texts in the corpus into registers based on their situational characteristics, and, in a number of cases, this required us to read through the texts to determine their primary communicative purposes. As shown in Table 1, the resulting corpus is both large (about twenty-million words) and represents a wide range of spoken and written registers.

3.3 Development of the grammatical tagger and identification of linguistic features

Before beginning work on grammatical analysis software, it was first necessary to identify the set of potentially relevant linguistic features to be used in the multi-dimensional analysis. For this purpose, we attempted to

itemise the linguistic characteristics of Spanish that potentially served communicative functions in discourse. We began by surveying major Spanish reference grammars, including the multi-volume *Gramática Descriptiva de la Lengua Española* (Bosque and Demonte, 1999), and various reference grammars written in English (especially including Butt and Benjamin, 2000). In addition, we consulted with Spanish grammarians and other native speakers. Finally, we considered the sets of linguistic features included in the MD analyses of other languages (especially English, Somali, and Korean) to check whether any of these had counterparts in Spanish.

Spoken texts: **Register**	**No. of texts**	**Word count**	**Average words/text**
Face–to-face conversations	111	259,568	2,338
Business telephone conversations	16	22,416	1,401
Sociolinguistic interviews	419	2,293,918	5,474
Political interviews	753	1,181,198	1,569
Radio/TV contests	23	53,813	2,340
Political debates	39	86,277	2,212
Drama	54	389,177	7,207
Institutional meetings	52	455,517	8,760
Political speeches	42	311,060	7,406
News broadcasts	31	68,309	2,204
Sports broadcasts	20	50,406	2,520
Spoken subtotal	1,560	5,171,659	3,315

Written texts: **Register**	**No. of texts**	**Word count**	**Average words/text**
Business letters	313	56,075	179
Fiction	187	7,205,389	38,531
Newspaper reportage	791	1,515,911	1,916
Editorials	49	95,603	1,951
Essays/Newspaper columns	378	1,977,167	5,231
General prose and textbooks	26	1,814,801	69,800
Encyclopaedias	708	2,304,457	3,255
Academic articles	40	160,785	4,020
Written subtotal	2,489	15,130,188	6,079

Table 1: Composition of the corpus used for the MD analysis

The grammatical tagger for our project, developed by Jones at Northern Arizona University, has several different components, including:

i) A probabilistic/rule-based component to identify the major word classes (nouns, verbs, adjectives, adverbs) together with basic morphological features (e.g., number, gender, tense),
ii) A rule-based component and morphological analyser to identify function word classes (e.g., prepositions, articles) and words belonging to special word classes (e.g., diminutives, nominalisations), and
iii) Rule-based components to identify additional semantic and syntactic features (e.g., semantic classes of verbs, complementation patterns, pro-drop).

The tagger was tested and revised extensively, checking the full set of features in texts from various registers, and then focusing on the especially problematic features (e.g., noun/verb/adj. ambiguities, and distinguishing among the functions of words like *se* and *que*). The overall accuracy of the final version of the tagger was estimated at 98 percent.

The tagger identifies about 140 different linguistic features. However, these were reduced to eighty-five features included in the final factor analysis (see section 3.4, below), as follows:

Vocabulary distributions: 1. type/token ratio, 2. average word length

Noun classes: 3. simple NPs (without articles, determiners, or numbers), 4. singular nouns, 5. plural nouns, 6. derived nouns (e.g., *-azo, -ión, -miento*), 7. proper nouns, 8. diminutives (*-ito*), 9. augmentatives (*-simo*)

Pronoun classes: 10. first person pronouns, 11. second person *tu* pronouns, 12. second person *usted* pronouns, 13. first person pro-drop, 14. second person pro-drop, 15. all third person pronouns except *se*, 16. reflexive *se*, 17. emoción *se*, 18. other *se* (not passive, reflexive, or "emoción"), 19. *conmigo/contigo/consigo*, 20. all clitics, 21. demonstrative pronouns (e.g., *ése*)

Adjective classes: 22. premodifying attributive adjectives, 23. postmodifying attributive adjectives, 24. predicative adjectives, 25. evaluative adjectives, 26. other semantic classes of adjective (colour, size/quantity/extent, time, classificational, topical), 27. quantifiers (e.g., *muchos, varias, cada*)

Other noun phrase elements: 28. definite articles, 29. premodifying demonstratives (e.g., *ese*), 30. possessives (including premodifying

determiners; pronouns, e.g., *la mía*; and emphatic pronouns, e.g., *hija mía*)

Adverb classes: 31. adverbs – place, 32. adverbs – time, 33. adverbs – manner, 34. other *-mente* adverbs

VERBS:

Tense and mood: 35. indicative, 36. subjunctive, 37. conditional, 38. present, 39. imperfect, 40. preterit, 41. progressive, 42. perfect, 43. future, 44. future time with *ir a*

Semantic/lexical classes: 45. obligation verbs (e.g., *deber, tener que, haber + que/de*), 46. all main verb *SER*, 47. All main verb *ESTAR*, 48. aspectual verbs, 49. mental and perceptual verbs, 50. verbs of desire, 51. communication verbs, 52. verbs of facilitation/causation, 53. verbs of simple occurrence, 54. verbs of existence/relationship

Other features of the verb phrase: 55. *ser* passive with *por*, 56. agentless *ser* passive, 57. *se* passive (with *por* and agentless), 58. verb + infinitive, 59. infinitives without preceding verb or article, 60. existential "haber"

Questions: 61. yes/no questions, 62. *CU* questions, 63. tag questions

Function word classes: 64. prepositions (single-word and multi-word), 65. general single-word conjunctions (*pero, y, e, o, u*), 66. other single-word conjuncts, 67. multi-word conjunctions, 68. exclamations (upside down exclamation mark)

DEPENDENT CLAUSES:

Adverbial clauses: 69. causal subordinate clause (e.g., *porque, puesto que, ya que*), 70. concessive subordinate clause (e.g., *aunque, a pesar de que*), 71. conditional clauses (e.g., *si, con tal que*)

Complement clauses: 72. *que* verb complement clause – indicative, 73. *que* verb complement clause – subjunctive, 74. *que* noun complement clause, 75. *que* adjective complement clause, 76. *CU* verb complement clause

Postnominal (relative) clauses: 77. *que* relative clause – indicative, 78. *que* relative clause – subjunctive, 79. *cual* relative clause, 80. *cuyo* relative clause, 81. postnominal past participles, 82. *el que* clauses

Other dependent clause features: 83. *que* clefts, 84. other *cual* clauses (not relatives), 85. conditional mood in dependent clause

After we completed the testing and revision of the tagger, we tagged the entire corpus. The following is an excerpt from a newspaper text, illustrating the various tag fields.

```
Pero ^con+coor+++++_gensingcon_+pero+
nada ^r+++++!!+_rbother_+nada+
de ^en++++++_1wrdprep_+de+
eso ^p3cs+dem+++++_prodem_+eso+
sucedió
^vm+is+3s++++_indicat_preter_voccur_+suceder+
, ^punc++,++++++
y ^con+coor+++++_gensingcon_+y+
el ^lms+def+++++_defart_+el+
embajador ^nms+com+++++_singn_derivn_+embajador+
concedió ^vm+is+3s++++_indicat_preter_+conceder+
su ^d3cs+pos+++++_prepos_+su+
mano ^nfs+com+++++_singn_+mano+
y ^con+coor+++++_gensingcon_+y+
la ^lfs+def+++++_defart_+la+
sonrisa ^nfs+com+++++_singn_+sonrisa+
imperturbable ^jcs++++++_postadj_+imperturbable+
a ^en++++++_1wrdprep_+a+
cada ^d0cs+ind+++++_quant_+cada+
uno ^p0ms+ind++++++uno+
de ^en++++++_1wrdprep_+de+
los ^lmp+def+++++_defart_+el+
invitados ^nmp+com++++!!+_plurn_+invitado+
```

Each line begins with the word followed by the start of the tag, indicated by ^. The primary tag is in field one (e.g., noun, verb, *etc.*), with various secondary tags in fields two to five (e.g., the mood, tense, person, number, and voice of a verb), an ambiguous tag in field six, a linguistic feature tag in field seven (e.g., "ynquest" for yes/no questions; "subjvcompque" for *que* verb complement clause with subjunctive mood), and the lemma in the final field.

Once the texts in the corpus were tagged, it was easy to compute the frequency of each linguistic feature in each text. These frequencies were "normalised" to a rate of occurrence per 1,000 words of text (see Biber, Conrad, Reppen, 1998: Methodology Box 6). Thus, at this stage, we had normed frequencies of eighty-five linguistic features for each text, making it possible to compute descriptive statistics for the different registers. The entire tagged corpus is available for research on the web, at http://www.corpusdelespanol.org/registers/. This site also includes tools for accessing the corpus, including searches on words, tags, or combinations of a word with tag sequences.

3.4 Factor analysis

As noted above, multi-dimensional analysis is a methodological approach that uses a statistical "factor analysis" to identify underlying patterns of linguistic co-occurrence. This procedure reduces a large number of original variables to a small set of underlying variables, called "factors". Each factor represents a group of variables that are correlated with one another (reflecting their statistical tendency to co-occur in texts); these factors can subsequently be interpreted as underlying "dimensions" of register variation.

The full factorial structure for the analysis of Spanish linguistic features can be seen in Appendix A. Only eighty-five of the original 140+ linguistic features were retained in the final factor analysis. Some features were dropped because they were redundant or overlapped to a large extent with other features. In other cases, features were dropped because they were generally rare in our corpus. Several of these features were combined into more general features. For example, possessive determiners and possessive pronouns were combined into a more general feature, *que* clefts includes both indicative and subjunctive clauses, and, similarly, *cual* relative clauses comprise a range of structural variants, including indicative and subjunctive clauses, with and without a preceding preposition. In addition, some features were dropped either because they did not vary across Spanish texts, or because they shared little variance with the overall factorial structure of this analysis (as shown by the communality estimates).

The solution for six factors was selected as optimal. These six factors account for 45 percent of the shared variance. A Promax rotation was used, which allows for some correlations of the factors. (Appendix A also shows the eigenvalues for the first six factors as well as the inter-factor correlations.)

Table 2 summarises the important linguistic features defining each dimension (i.e., features with factor loadings over + or – .3). Each factor comprises a set of linguistic features that tend to co-occur in the texts from the Spanish corpus. Factors are interpreted as underlying "dimensions" of variation based on the assumption that linguistic co-occurrence patterns reflect underlying communicative functions. That is, particular sets of linguistic features co-occur frequently in texts because they serve related communicative functions. Features with positive and negative loadings represent two distinct sets of co-occurrence. These define a single factor because the two sets tend to occur in complementary distribution: when a text has a high frequency of the positive set of features, that same text will tend to have low frequencies of the negative set of features, and *vice versa*. In the interpretation of a factor, it is important to consider, (i) the communicative functions that are shared by the linguistic features grouped on a dimension, (ii) the patterns of register variation with respect to the group of linguistic features, and (iii) the functions of target linguistic

features in particular texts. In the following section, we present the interpretations of each factor as a dimension of variation.

4. Interpretation of the Spanish dimensions of variation

4.1 Interpretation of Dimension 1: 'Oral' versus 'literate' discourse

The first step in the interpretation of a dimension is to describe the communicative functions shared by the co-occurring linguistic features. In the case of Dimension 1, there is an extremely large number of linguistic features with large positive weights, and these features can be described as serving several specific functions. However, those functions are related in that they are all characteristic of spoken language rather than written language.

Many of these features are verb classes or characteristics of the verb phrase, such as indicative mood, present tense, future *ir a*, perfect aspect, and progressive aspect. (By contrast, there are almost no nominal features included in the positive grouping on Dimension 1.) Several of these verbal features are used for simple descriptions, including copula *SER*, copula *ESTAR*, existential *haber*, and simple occurrence verbs (e.g., *pasar, occurrir*). Some of these verb classes – especially mental verbs, desire verbs, and the copula *ESTAR* – frequently occur with first person pronouns (and first person pro-drop) to express the speaker's own personal feelings and attitudes. There are also several 'addressee-oriented' features included on Dimension 1, such as the pronouns *tú* and *usted*, tag questions, and yes-no questions. And at the same time, there are several 'other-oriented' features grouped on to this dimension, reflecting the description of other people in particular places and times (e.g., features like third person pronouns, time adverbs, place adverbs, demonstrative pronouns, communication verbs, and manner adverbs).

These positive features can all be associated with stereotypical "oral" discourse, and this interpretation is supported by the patterns of register variation along Dimension 1 (described below). However, given that interpretation, it might surprise many readers that there are also several dependent clause types grouped with the positive Dimension 1 features: causal subordinate clauses, conditional subordinate clauses, *que* verb complement clauses (indicative), *CU* verb complement clauses, *el que* clauses, and *que* relative clauses (indicative). Similar patterns have been found in the MD analyses of other languages, where adverbial and complement clauses are commonly used to express personal stance and thus they co-occur with features like pronouns and verbs in spoken discourse.

Dimension 1

Positive features:

indicative mood, causal subordinate clauses, time adverbs, first person pronouns, copula *SER*, demonstrative pronouns, specific single-word conjuncts, first person pro-drop, copula *ESTAR*, mental verbs, place adverbs, existential *haber*, *que* verb complement clauses (indicative), tag questions, present tense, future *ir a*, perfect aspect, communication verbs, third person pronouns, progressive aspect, *el que* clauses, yes-no questions, *que* relative clauses (indic.), manner adverbs, augmentatives, quantifiers, *CU* verb complement clauses, premodifying demonstratives, conditional subordinate clauses, *tu*, *usted*, desire verbs, general single-word conjuncts, verbs of facilitation, simple occurrence verbs

Negative features:

singular nouns, postmodifying adjectives, definite articles, prepositions, plural nouns, simple NPs (without determiners, *etc.*), derived nouns, type token ratio, postnominal past participles, premodifying attributive adjectives, long words, other adjectives, *se* passives

Dimension 2

Positive features:

subjunctive verbs, *que* relative clauses (subjunctive), *que* verb complement clauses (subjunctive), verb+infinitive, conditional verbs, obligation verbs, future tense, infinitives without preceding verb or article, *que* verb complement clauses (indicative), verbs of facilitation, progressive aspect, conditionals in dependent clauses, *que* noun complement clauses

Dimension 3

Positive features:

clitics, imperfect tense, possessives, third person pronouns, *se* (not passive or reflexive), preterit tense, aspectual verbs, *se* (reflexive), *se* (*emoción*), infinitives without preceding verb or article, verb+infinitive

Negative features:

derived nouns, postmodifying adjectives

Dimension 4

Positive features:

third person pro-drop, *tu*, exclamatives, *CU* questions, simple NPs (without determiners, *etc.*), yes-no questions, diminutives, *conmigo/contigo/consigo*

Negative features:

que relative clauses (indicative), other *-mente* adverbs

Dimension 5

Positive features:

proper nouns, preterit tense, long words, prepositions, premodifying attributive adjectives

Negative features:

present tense, predicative adjectives, verb+infinitive

Dimension 6

Positive features:

cual relative clauses, other *cual* clauses

Table 2 (previous page): Summary of the important linguistic features defining each dimension

At the other extreme, the negative features grouped on Dimension 1 are mostly nominal – types of nouns or characteristics of noun phrases. These include singular nouns, postmodifying adjectives, definite articles, plural nouns, simple NPs (without determiners, *etc.*), derived nouns, postnominal past participles, and premodifying attributive adjectives. Prepositional phrases always contain a noun phrase, and they often function to modify some head noun. Long words and a high type token ratio are also included with these negative features, reflecting the use of a diversified vocabulary and specialised words. The reliance on nouns and complex noun phrases results in a style of text with dense informational content packed into relatively few words. Writers, who have extensive opportunity to craft and revise their texts, are able to achieve this linguistic style of expression, but it is rare to find spoken texts of this type. Thus, Dimension 1 can be interpreted as reflecting the characteristics of stereotypical "oral" discourse (the positive features) versus "literate" discourse (the negative features).

This interpretation is strongly supported by the patterns of register variation found with respect to Dimension 1 (see Figure 1). That is, the second major step in interpreting a dimension is to consider the similarities and differences among registers with respect to the set of co-occurring linguistic features. For that analysis, *dimension scores* are computed for each text, and then texts and registers are compared with respect to those scores. Dimension scores (or *factor scores*) are computed by summing the individual scores of the features with salient loadings on a dimension (i.e., features with loadings greater then |.30| on a factor).

In the present case, the Dimension 1 score for each text is computed by adding together the frequencies of indicative mood verbs, verbs of existence, causal subordinate clauses, time adverbs, first person pronouns, copula *SER*, *etc.* – the features with positive loadings on Factor 1 (from Table 2) – and then subtracting the frequencies of singular nouns, postmodifying adjectives, definite articles, prepositions, plural nouns, *etc.* – the features with negative loadings.

All individual linguistic variables are standardised to a mean of 0.0 and a standard deviation of 1.0 before the dimension scores are computed. This process converts feature scores to scales representing standard deviation units, so that all features on a factor have equivalent weights in the computation of dimension scores (see Biber, 1988: 93-97).

Once a dimension score is computed for each text, the mean dimension score for each register can be computed. Plots of these mean dimension scores allow linguistic characterisation of any given register, comparison of the relations between any two registers, and a fuller functional interpretation of the underlying dimension.

Figure 1 shows the mean dimension scores of registers along Dimension 1. The registers with large positive values (such as telephone and casual face-to-face conversations), have high frequencies of indicative mood verbs, verbs of existence, first person pronouns, *etc.* – the features with salient positive weights on Dimension 1. At the same time, these registers with large positive values have markedly low frequencies of singular nouns, postmodifying adjectives, prepositions, *etc.* – the features with salient negative weights on Dimension 1. Registers with large negative values (such as academic prose and encyclopaedias) have the opposite linguistic characteristics: very high frequencies of nouns, postmodifying adjectives, prepositions, *etc.*, plus low frequencies of verbs, pronouns, and so on.

The register distribution shown in Figure 1 confirms the interpretation of Dimension 1 as a continuum of "oral" versus "literate" discourse. In fact, there is almost an absolute distinction between spoken versus written registers along Dimension 1: all spoken registers have positive scores on Dimension 1, while all written registers – with the exception of fiction – have negative scores on Dimension 1. Within speech, the conversational registers have the largest positive scores; these

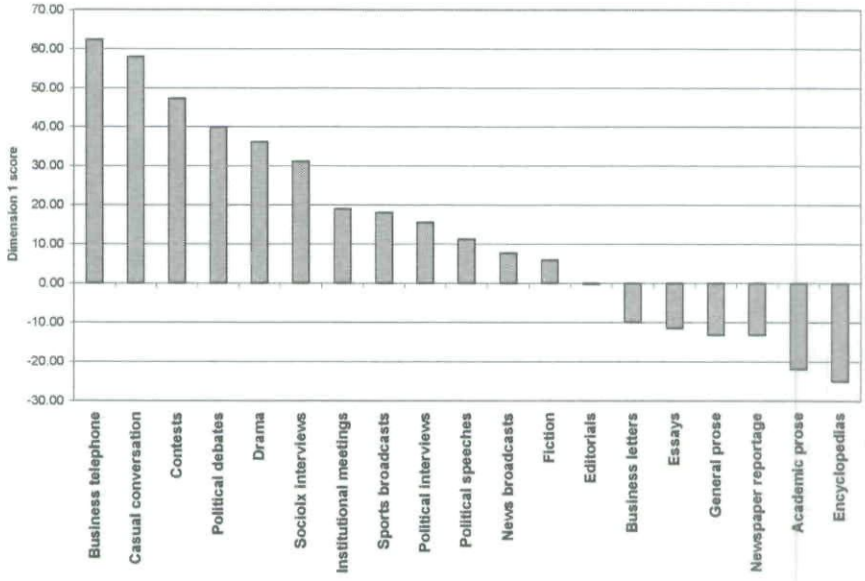

Figure 1: Comparison of registers along Dimension 1: "oral" versus "literate" discourse

registers are highly interactive and involved, but the language is minimally planned ahead of time. For example, Text Sample 1 is from a casual face-to-face conversation. This dialogue includes several references to both the

speaker and addressee, as well as a short narrative about a third person. Overall, the interaction relies heavily on verbs and short clauses.

Text Sample 1: Casual conversation
(Present tense verbs are underlined. Most of those verbs are also examples of cases where the subject pronoun is dropped.)

Speaker 1: Pero, ¿ese Alberto es - es un alumno tuyo - extranjero?
Speaker 2: Sí.
Speaker 1: ¿Y de dónde es, llamándose Alberto?
Speaker 2: Italiano.
Speaker 1: Ah, italiano. Claro, claro.
Speaker 2: Di/ - tiene, fíjate, se llama Alberto y tiene un apellido catalán.
Speaker 1: ¿Y eso?
Speaker 2: ¿Eh?
Speaker 1: ¿Y eso?
Speaker 2: Pues no sé. Digo: "¡Pero bueno!" Dice: "Sí, sí", dice: "Si - muchas veces me he hecho pasar por español". [...]

Translation:

Speaker 1: But, ¿This Alberto is - is a student of yours - a foreigner?
Speaker 2: yes
Speaker 1: ¿And where is he from, with a name like Alberto?
Speaker 2: Italian
Speaker 1: Ah, italian. Sure, sure.
Speaker 2: He has, get this, his name is Alberto and he has a Catalan last name.
Speaker 1: ¿And so?
Speaker 2: ¿huh?
Speaker 1: ¿And so?
Speaker 2: So, I don't know. I say: "¡Well, good!" He says: "yes," yes", he says: "yes - a lot of times I've gotten others to think I'm Spanish". [...]

At the other end of the spoken continuum, registers like institutional meetings, political interviews, political speeches, and news broadcasts are planned and much less directly interactive than conversation. At the same time, these registers are much more focused on conveying

information than conversation. As a result, spoken registers like political speeches and news broadcasts have small positive scores, reflecting a more balanced use of positive and negative features on Dimension 1.

As noted above, the written registers (except fiction) all have negative scores on Dimension 1, with academic prose and encyclopaedias having the largest negative scores. These scores reflect the dense use of nominal features (e.g., nouns, postmodifying adjectives, prepositions, *etc.*) together with the infrequent use of positive Dimension 1 features (verbs, pronouns, *etc.*). Text Sample 2 illustrates the use of these features in an academic prose text:

Text Sample 2: Academic prose
(Nouns are underlined.)

También es común en estas zonas la desaparición de corrientes de agua, e incluso ríos enteros pueden desaparecer en sumideros, u ojos que pueden conducir a cavernas subterráneas o a acuíferos. Los sumideros indican la presencia de cuevas bajo ellos. Debido a la captura de las aguas superficiales por el sistema subterráneo de drenaje, algunas regiones con cuevas son bastante secas y polvorientas y tienen escasa vegetación.

Translation:

Underground streams are also common, and even entire rivers can sometimes disappear in sinkholes that lead to underground caves or aquifers. The sinkholes indicate the presence of caves underneath. Due to the flow of surface-level water into an underground system of drainage, some areas with many caves are rather dry and have little vegetation.

In sum, Dimension 1 makes a fundamental distinction between speech and writing, and, at the two extremes, this dimension actually distinguishes between stereotypical speaking (conversation) and stereotypical writing (expository prose). Linguistically, these opposing styles are represented by verbal/clausal features serving involved and interactive functions, versus a dense nominal packaging of information that requires careful production and revision of the text itself.

4.2 Interpretation of Dimension 2: spoken 'irrealis' discourse

Only positive features are grouped on Dimension 2, and these mostly relate to the expression of opinions and the description of hypothetical situations.

These constructions describe personal feelings and attitudes, or possible events/states, but they do not describe an *actual* event or state. Several of the linguistic features grouped on this dimension include a subjunctive verb, which is used for the expression of possibility/probability, desire, persuasion, doubt, fear, hope, *etc.* (see Butt and Benjamin, 2000: 238 ff.). Conditional verbs and the future tense are both used to describe events or states that *could* occur, but have not actually occurred. Similarly, obligation verbs (e.g., *tengo que*) describe events that should occur. Finally, *que* verb complement clauses and *que* noun complement clauses are used to express a stance in the controlling verb or noun (e.g., *sabe que*, *la idea de que*).

The register differences defined by Dimension 2 are in many ways similar to those found on Dimension 1. Figure 2 shows that Dimension 2 defines a near absolute distinction between spoken and written registers: only spoken registers have large positive scores on Dimension 2 (reflecting the dense use of these 'irrealis' features). By contrast, all written registers have negative scores or scores near 0.0; and the two formal expository written registers – encyclopaedias and academic prose – have large negative scores, reflecting the marked absence of these irrealis features.

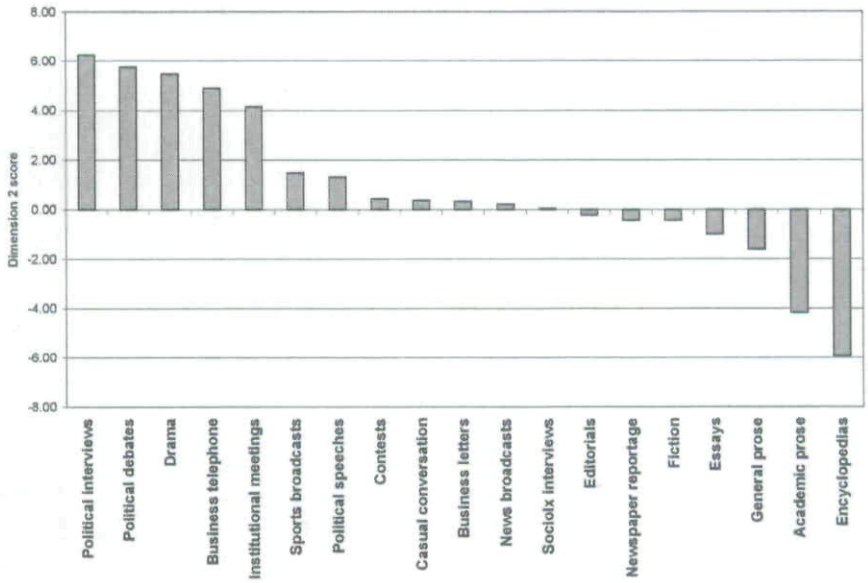

Figure 2: Comparison of registers along Dimension 2: spoken irrealis discourse

There are, however, important differences between the register patterns of Dimensions 1 and 2. As noted above, Dimension 1 reflects stereotypical speaking versus writing, so that the conversational registers

are especially marked with large positive scores. By contrast, on Dimension 2 casual conversation has a score near 0.0. Instead, we see the opinionated spoken registers with large positive scores on Dimension 2: especially political interviews and political debates. For example:

Text sample 3: Political interview
(Subjunctive, future, and conditional verbs are underlined.)

Speaker 1: ...Hay que ser fuerte pero a la vez generoso, dispuestos a tratar al otro con respeto y compasión.
Speaker 2: Una actitud moral -¿Y servirá de algo la música?-
Speaker 1: Servirá solamente si la gente llega a saber como amarla, como cantar e improvisar. Escuchar música en un sillón está muy bien. Claro está que me gusta un público que sepa escuchar. Pero en realidad me gustaría que la música y el canto fueran siempre previos a los debates políticos. Enseñar música en las escuelas con la actitud moral adecuada podría contribuir al entendimiento en el mundo.

Translation:

Speaker 1: ...You have to be strong but at the same time generous, prepared to treat the other with respect and compassion.
Speaker 2: A moral attitude - ¿And would music help? -
Speaker 1: It would help only if people really learn to enjoy it, how to sing (along). Listening to music in a big, soft chair is a good idea. You know, I really like people who know how to listen. But actually, I'd really like music and singing to be part of the public dialgoue. Teaching music in the schools with the right kind of attitude would really contribute to understanding in the world.

Interestingly, drama and formal telephone conversations also show a dense use of these features, whilst casual face-to-face conversation does not. In the case of drama, the dialogue carries the narrative storyline while

showing us the character's inner thoughts and feelings, resulting in a dense use of these irrealis features.

Surprisingly, no written register is marked for the dense use of these features. That is, even registers like editorials and essays are characterised by the relative absence of irrealis features, despite their communicative goals of arguing for a particular point of view in opposition to other possible perspectives. In part, this is due to the fact that editorials and essays (at least those found in our corpus) often have a past-tense "narrative" orientation. This allows the columnist to relate a series of events that deal with the overall argument that he or she is making in the column. These narrative passages have a lower degree of "irrealis" than the present and future-orientated debates and drama.

4.3 Interpretation of Dimension 3: narrative discourse

The positive features on Dimension 3 are commonly used to construct stereotypical narrative discourse. Imperfect and preterit tense verbs form the backbone of this discourse, presenting events and background descriptions in past time. Third person pronouns, possessives, and clitics are similarly important to refer to the participants in the narrative.

In developing the grammatical tagger for this project, we made a considerable effort to distinguish between the uses of the particle *se*, including reflexive verbs (e.g., *lavarse*, 'to wash oneself'), the *se - emocion*

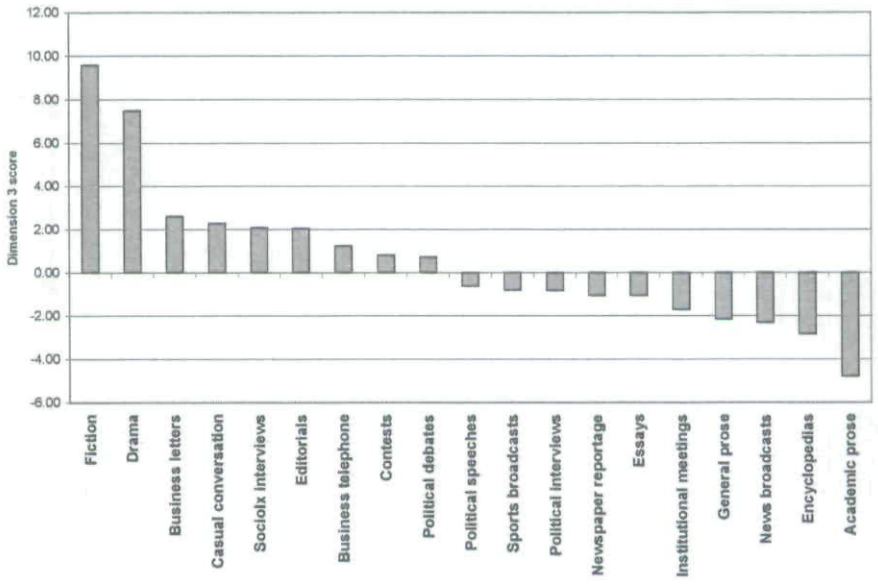

Figure 3: Comparison of registers along Dimension 3: narrative discourse

(e.g., *alegarse*, 'to be happy'), the passive *se*, and other uses (e.g., *comerse*, 'to eat up'). However, the factor analysis grouped together all these uses of *se* – except for the passive *se* – on Dimension 3, showing that they commonly co-occur in narrative texts (apparently because they all function as a focusing element allowing foregrounding).

As Figure 3 shows, fiction has by far the largest positive score on Dimension 3. The following text sample illustrates the dense use of Dimension 3 features in a fiction text, including both imperfect and preterit verbs (*tenia, vieron*), clitics (*hablarles*), possessives (*su*), and verbs with *se*.

Text Sample 4: Fiction
(Third person pronouns, possessives, and clitics are underlined.)

Aquella noche, al dar la hora de acostarse, don José Pedro llamó a sus dos hijas.
-Su madre quiere hablarles. En su pieza las espera -les dijo.
Y fumando, bajó al parque. Las hijas lo vieron perderse por entre la sombra
densa. Tan oscura estaba la noche, que todo tenía el color de las araucarias...
Las dos acudieron al cuarto de su madre.
-Siéntense. Quiero verlas muy serenas en este momento -empezó la señora-.
¿Se ha tranquilizado ya Demetrio? Por lo que observé durante la comida, todo malentendido pasó. Bien. Ahora; juntas, lean esta carta. Juntas y en silencio.

Translation:

That night, around bedtime, Don Jose Pedro called together his two daughters. "Your mother wants to talk to you. She's waiting for you in the bedroom" he said. And so, smoking, he left for the park. The daughters saw him disappear into the dense darkness. It was so dark outside that everything was the color of dark pine trees.
The two girls went straight into their mother's room.
"Sit down. I want you to settle down now" the woman began. "Has Demetrio settled down now as well?" From what I saw at dinner, no one has understood anything. Okay. Now I want you to read this letter - both together and in silence."

Interestingly, drama also has a large positive score on Dimension 3. In this case, the text is entirely dialogic, but the characters are narrating past events and descriptions to carry the storyline of the play. In addition to

fiction and drama, several other registers have moderate positive scores on Dimension 3, showing that, to some extent, they incorporate narratives in their discourse. These include both spoken and written registers, such as casual conversations and sociolinguistic interviews (spoken), and business letters and editorials (written).

At the other extreme, academic prose (and to a lesser extent encyclopaedia articles) are marked by the absence of these narrative features. Texts from these registers are normally descriptive or explanatory, rather than narrating past events, and so they are characterised by the absence of the positive Dimension 3 features. Text Sample 2 (above) illustrates written prose of this type.

4.4 Interpretation of Dimension 4: addressee-focused interaction

Dimension 4 is composed of overtly interactive and highly involved features, including *CU* questions, yes-no questions, exclamatives, and diminutives. However, by contrast with Dimension 1, the style of discourse presented here seems to be focused to a large extent on the addressee, resulting in the dense use of second person pro-drop, and the pronoun *tu*, but not first or third person pronouns.

This somewhat specialised grouping of linguistic features is especially common in business telephone conversations (see Figure 4). In this register, telephone operators are interacting with customers, obtaining

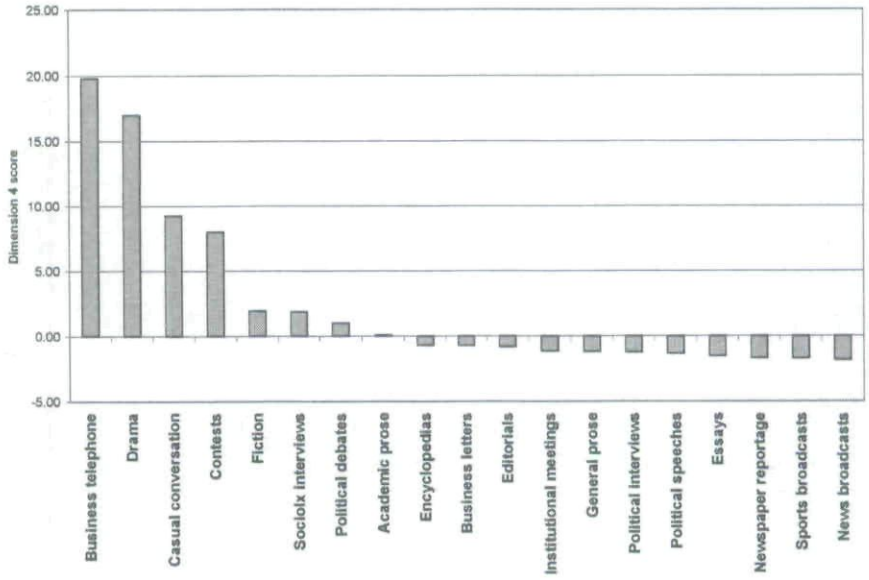

Figure 4: Comparison of registers along Dimension 4: addressee-focused interaction

information and attempting to help with customer problems. In our corpus, these are conventionalised interactions that focus on the addressee, with little expression of the feelings and attitudes of the speaker, as shown in the following example.

Text Sample 5: Business telephone conversation
(Questions begin and end with ¿ ?)

Speaker 1:	Perdóname un segundito. "Cilag", ¿dígame?
Speaker 2:	¿Teresa?
Speaker 1:	Sí.
Speaker 2:	Hola, soy Miguel.
Speaker 1:	Hola, Miguel. ¿Qué te cuentas?
Speaker 2:	¿Qué tal, cómo estás?
Speaker 1:	Dime.
Speaker 2:	Sí, ¿me pasas con Rocío?
Speaker 1:	Te pongo con ella.
Speaker 2:	Vale.
Speaker 1:	¿Sí?
Speaker 3:	Miguel Llavori.
Speaker 1:	Vale.
Speaker 3:	¿Marica?
Speaker 1:	Sí.
Speaker 3:	Sí, eh - ¿qué empiezas, por lo de Tenerife?
Speaker 1:	Sí, te cuento lo de Tenerife.

Translation:

Speaker 1:	OK. Excuse me. "Cilag". Hello?
Speaker 2:	Teresa?
Speaker 1:	Yes.
Speaker 2:	Hi, it's Miguel.
Speaker 1:	Hi Miguel. What's happening?
Speaker 2:	So, how's everything going? How are you?
Speaker 1:	So…?
Speaker 2:	Look, can I talk to Rocío?
Speaker 1:	Sure, I'll connect you.
Speaker 2:	Thanks.
Speaker 1:	Hello?
Speaker 3:	(Hi, it's) Miguel Llavori.
Speaker 1:	Yeah?
Speaker 3:	Marica?
Speaker 1:	Yeah?
Speaker 3:	Well, um, can you tell me about the Tenerife deal?
Speaker 1:	Yeah, let me tell you.

4.5 Interpretation of Dimension 5: informational reports of past events

In common with Dimension 3, the features grouped on Dimension 5 also relate to the reporting of past events. However, there are important differences between these two dimensions. There are a relatively large number of linguistic features grouped on Dimension 3, including imperfect and preterit tense, various clitics, and third person pronouns. As we saw above, these features are especially common in fictional narrative and drama. By contrast, Dimension 5 is much more specialised. It is defined by a smaller set of features: proper nouns, preterit tense, long words, prepositions, and premodifying attributive adjectives. (The major negative features are present tense, predicative adjectives, and verb+infinitive.)

Although the positive features grouped on Dimension 5 are related to past time discourse, they are quite different from the Dimension 3 features. First of all, Dimension 5 includes only preterit tense (but not imperfect tense verbs), reflecting a focus on past time events with relatively little background description. In addition, we find proper nouns with a large positive loading on Dimension 5, rather than third person pronouns. This suggests a style of discourse that focuses on the past actions of many different people, referred to by name. By contrast, Dimension 3 features characterise more detailed fictional narratives that involve a few characters, which are easily referred to with third person pronouns. Furthermore, Dimension 5 includes features of highly informational prose – long words, prepositions, and premodifying attributive adjectives – suggesting that this style of discourse has an informational, rather than popular, communicative purpose.

As Figure 5 shows, these features are common only in written informational registers, that is, encyclopaedias, business letters, newspaper reportage, and, to a lesser extent, academic prose. Encyclopaedias and newspaper reportage are similar in that they are informational registers written for a mass audience, informing readers about past events that involve many different people. Text sample 6, from an encyclopaedia article, illustrates these features:

Text Sample 6: Encyclopaedia article
(Preterit verbs and proper nouns are underlined.)

Tras abandonar los terrenos de juego, Suárez inició una nueva carrera como técnico. En esta faceta profesional, permaneció casi siempre ligado a la secretaría técnica del Inter de Milán, de cuyo primer equipo llegó a ser entrenador. También ocupó el banquillo de varios clubes españoles. Además, estuvo al frente de la selección nacional absoluta española de fútbol, a la cual dirigió en la fase final de la Copa del Mundo disputada en 1990 en Italia. En 1992 regresó al Inter, primeramente como

entrenador y, más tarde, como integrante de su equipo técnico.

Translation:

After retiring, Suárez began a new career as a technical advisor. In this professional capacity, he was associated with the technical staff of Inter in Milan, and he became trainer for their first-string team. He was also an advisor for several Spanish teams. In addition, he was in charge of the selection of the national team for Spain, which he led to the final round of the World Cup in 1990 in Italy. In 1992, he returned to Inter, first as a trainer and then as one of the principal members of their technical staff.

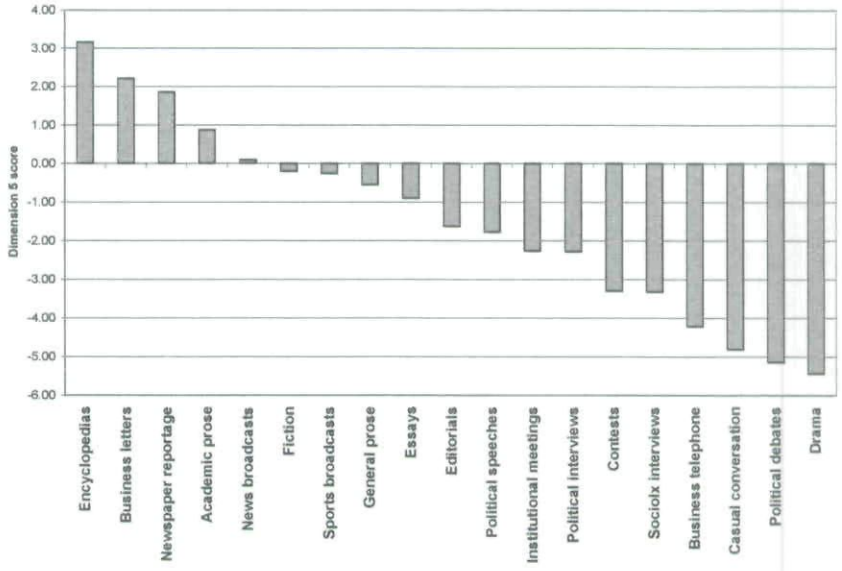

Figure 5: Comparison of registers along Dimension 5: informational reports of past events

The distinction between Dimensions 3 and 5 reflects the differing functions of the imperfect versus the preterit, two forms of the past tense in Spanish that differ in aspect (a distinction not similarly found in English). The preterit refers strictly to events that are viewed as a single whole, and preterit verbs would therefore be common in encyclopaedias and news reports (with the highest positive scores on Dimension 5). The imperfect, on the other hand, describes an event that was not yet complete, and thus it is used for background descriptions of events that were in progress or states

that existed when another event occurred. These discourse functions are important for the description and narration typical of drama and fiction, the registers with the largest positive Dimension 3 scores. It is interesting that the multi-dimensional structure reflects this grammatical distinction found in Spanish (but not English), although we return to this point in the conclusion.

4.6 Interpretation of dimension 6: 'formal' written style

Finally, Dimension 6 is an extremely specialised parameter defined by only two co-occurring linguistic features: *cual* relative clauses and other *cual* clauses. As Figure 6 shows, these features are common only in formal written prose, and especially in academic prose. We tentatively interpret this dimension as reflecting a formal "high" academic style of discourse.

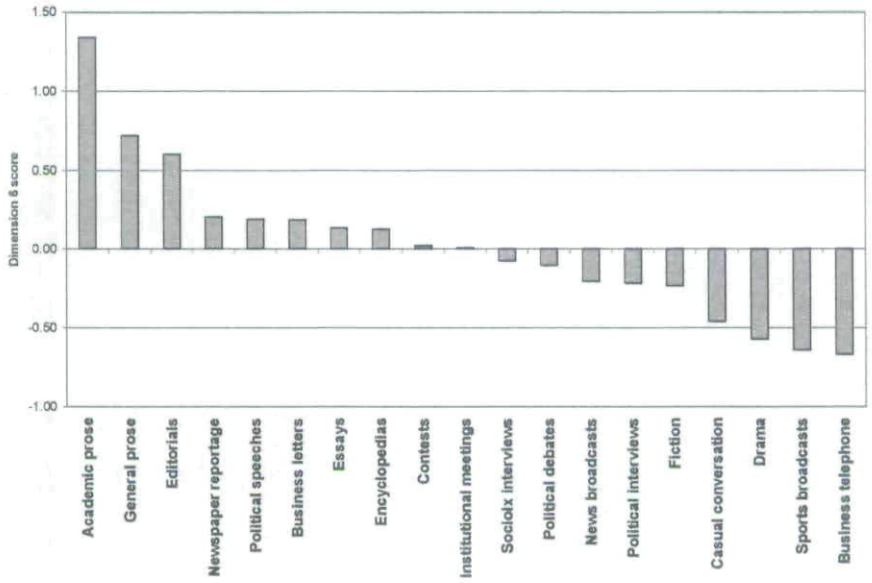

Figure 6: Comparison of registers along Dimension 6: 'formal' written style

5. Discussion and conclusion

There have been previous MD studies of register variation in several languages, including Biber's (1985, 1986, 1988) analysis of English, Besnier's (1988) analysis of Nukulaelae Tuvaluan, Kim's (1990; Kim and Biber, 1994) analysis of Korean, Biber and Hared's (1992a, 1992b, 1994) analysis of Somali, and Jang's (1998) study of Taiwanese. Biber (1995)

synthesises these studies, describing striking similarities in the basic patterns of register variation, as reflected by:

- the co-occurring linguistic features that define the dimensions of variation in each language;
- the functional considerations represented by those dimensions; and,
- the linguistic/functional relations among analogous registers.

Probably the most striking similarity across languages is that the first dimension in each case defines a basic opposition between oral and literate registers. These dimensions are similar in their linguistic composition and in the register differences that they define. For most languages, the positive linguistic features on these dimensions include interactive features, reduced structure features, and stance features. By contrast, the negative linguistic features grouped on the first dimension in each of these languages include noun features, adjectival features, and noun modifiers. For all languages, conversational registers are at the extreme positive pole of the first dimensions, whilst written expository registers are at the negative pole. Functionally, these dimensions are interpreted as reflecting direct interaction, real-time production circumstances, and personal stance and involvement, versus an informational focus and registers that permit carefully planned and revised production. The fact that this dimension emerges as the first factor in the analyses of all these languages suggests that it represents a fundamentally important parameter of register variation across languages.

A second similarity across all MD analyses is the existence of a narrative dimension. In all cases, this dimension consists of linguistic features associated with stereotypical narrative discourse, such as past tense verbs, third person pronouns, and temporal adverbials. Written fiction is consistently the most marked register on this dimension, but spoken folktales also have high positive scores. In several of the languages, certain kinds of public speaking have intermediate scores on the narrative dimension, reflecting the incorporation of narratives into public speeches or sermons.

At the same time, each of these MD analyses has identified dimensions that are unique to a language, reflecting the particular communicative priorities of that language and culture. For example, the MD analysis of Somali identified a dimension interpreted as 'Distanced, directive interaction' (for example, defined by optative clauses, first and second person pronouns, and directional particles). Only one register is especially marked for the frequent use of these features in Somali: personal letters. This dimension reflects the particular communicative priorities of personal letters in Somali, which are typically interactive but explicitly directive.

The Spanish MD analysis offers further evidence for both of these two major patterns: the existence of cross-linguistic universals, together with distinctive dimensions associated with each language/culture. Probably the most noticeable similarity between the Spanish MD analysis and previous analyses is Dimension 1. For example, the positive linguistic features on Spanish Dimension 1 include interactive features and stance features – mostly verb classes and clausal features. By contrast, the negative features on Dimension 1 are mostly phrasal, and associated with nouns and noun phrase modification (nouns, adjectives, definite articles, prepositions, *etc.*). This dimension distinguishes between oral and literate registers, and it is interpreted functionally in relation to interactiveness, personal involvement, and stance, by contrast with the primary informational focus of expository writing. In all these respects, Dimension 1 in the Spanish MD analysis is strikingly similar to the first dimension in previous MD analyses.

A second major point of similarity is the existence of a narrative discourse dimension, that is, Spanish Dimension 3, which consists of past tense verbs, third person pronouns, and the particle *se*. As with the narrative dimension in other languages, this dimension distinguishes between narrative fiction (and drama) versus expository written registers.

A third point of similarity is more subtle and surprising: the MD analyses of all languages have shown that certain kinds of structural complexity are associated with speech rather than writing. In particular, two kinds of dependent clause – complement clauses (especially controlled by verbs) and adverbial clauses – are consistently grouped together with the features of high interaction and personal involvement, among the positive features of Dimension 1. The MD analysis of Spanish further supports this association, and the positive features on this dimension include *que* verb complement clauses, *CU* verb complement clauses, causal adverbial clauses, and conditional adverbial clauses.

At the same time, there are major differences between the MD analysis of Spanish and the analyses of other languages. Two of these are especially noteworthy: the existence of two 'past time' dimensions in Spanish, and the existence of a 'spoken irrealis' dimension.

The first difference relates to the existence of two distinct 'past time' dimensions in Spanish: Dimensions 3 and 5. Dimension 3 is very similar to the narrative dimension identified in previous MD analyses, consisting of past tense verbs (both preterit and imperfect) and third person pronouns. Written fiction is the most marked register along this dimension. By contrast, Dimension 5 is distinctive, and unlike any dimension identified in previous MD analyses. Dimension 5 consists of only one of the two tenses that express past time in Spanish – the preterit – co-occurring with nominal features associated with an informational focus (proper nouns, long words, prepositions, attributive adjectives). This dimension has a more specialised function in Spanish, distinguishing between expository registers that have a primary informational focus on

reporting past events (such as encyclopaedias and newspaper reportage), versus all other spoken and written registers. Interestingly, fiction has a slightly negative score on Dimension 5.

The second distinctive Spanish dimension is the 'spoken irrealis' dimension (D2). The features on Dimension 2 are mostly used for the expression of opinions and the description of hypothetical situations, describing personal feelings and attitudes, or possible events/states, rather than an actual event or situation (e.g., subjunctives, conditional verbs, future tense, *que* complement clauses). This dimension is similar to Dimension 3 in the Korean MD analysis (see Biber, 1995: 193-96), which is also defined exclusively by stance features. These two dimensions are also similar in that they distinguish generally between spoken registers (marked for the dense use of stance features) versus written registers (marked for the absence of these features). However, the Spanish Dimension 2 is distinctive in two respects: first, it includes several features relating to irrealis discourse, in addition to epistemic and attitudinal stance features, and, secondly, it distinguishes between opinionated or persuasive spoken registers (for example, political interviews and political debates) and all other registers (including casual face-to-face conversation).

There are important possible confounding influences that must be considered when interpreting cross-linguistic MD comparisons. The most important consideration has to do with differences in the corpus design investigated for each language. For example, some MD analyses have been based on existing corpora, which can be limited in size and/or design, and the 1988 MD study of English was based on a one-million words sample from the LOB and London-Lund corpora. By contrast, Kim designed and constructed a corpus specifically for his 1990 MD study of Korean, but, due to limited resources, that corpus includes only about 200,000 words. Similarly, Biber and Hared designed and constructed a corpus specifically for the MD study of Somali. However, because their project was sponsored by a federal research grant (NSF), the corpus includes a more comprehensive set of spoken and written registers than we find in previous studies (see Biber, 1995: 90-93).

The present MD study of Spanish is based on a much larger corpus than any previous MD study (see Section 3.2). However, that corpus is compiled from a combination of pre-existing corpora, rather than being designed specifically for our study. As a result, the corpus is skewed in some respects relative to previous MD analyses. For example, face-to-face conversations constitute only 7 percent of the total texts in the spoken sub-corpus (111 out of a total of 1,560 texts), while the much more specialised registers of sociolinguistic interviews and political interviews have much larger representations (27 percent and 48 percent of the total texts). Similarly, academic research articles constitute only 1.6 percent of the total texts in the written sub-corpus (forty out of a total of 2,489 texts), while the more specialised register of encyclopaedia articles constitutes 28 percent of the total. Thus, when comparing the multi-dimensional structure of these

languages, it is important to bear in mind the possible influence of differences in corpus design. In general, however, that influence has been relatively minor, because the corpora all cover roughly the same range of registers, differing primarily in the relative weightings given to particular registers.

A second possible confounding influence for cross-linguistic comparisons is that each of these languages has a different inventory of structural devices and distinctions. The multi-dimensional patterns for each language reflect a complex interaction between the structural resources available in the language and the register distinctions that are systematically marked by those resources. For example, the existence of subjunctive mood verbs in Spanish provides the linguistic resources for a dimension associated with irrealis discourse. Similarly, the existence of two past tenses in Spanish provides the structural resources for a specialised dimension associated with informational reports of past events.

The existence of structural distinctions does not necessarily entail the existence of systematic register differences, but languages/cultures have often evolved to take advantage of these linguistic resources. Previous MD analyses identify several cases where specialised structural distinctions are systematically exploited to make specialised register distinctions. The existence of specialised dimensions relating to irrealis discourse and informational reports of past events in Spanish reflect this tendency. The Korean MD analysis provides another strong example of this tendency: personal stance features are grouped on one dimension, while features of honorification and self-humbling are grouped on a separate dimension.

Previous MD analyses show that the ways in which a language/culture exploits such structural resources are not always what we would have anticipated. For example, in Korean, the co-occuring features associated with the 'stance' dimension (for example, emphatics, hedges, other epistemic and attitudinal features) are especially common in the (inter)personal registers, including all conversations and personal letters. By contrast, the features associated with the honorific/self-humbling dimension are especially common only in the *public* spoken registers, such as public interviews and public speeches. Both dimensions are generally related to the expression of stance. However, the MD analysis shows that they are exploited in different ways for specific cultural purposes.

The MD analysis of Spanish has similar unanticipated findings. For example, it is perhaps not surprising that Spanish-speaking cultures would evolve to exploit subjunctive mood verbs for irrealis purposes. However, it is more surprising that these features tend to co-occur with a range of other stance features, and that they are used primarily in *spoken* opinionated registers, (whilst they are relatively rare in written opinionated registers).

These patterns illustrate the general finding that structural resources come to be exploited in particular (often unanticipated) ways in particular cultures. Some linguistic features are distributed widely across

different languages, and they are exploited in very similar ways to distinguish between registers across cultures. For example, features like first and second person pronouns, questions, reduced/contracted forms, and simple hedging or emphatic stance features are found in many languages, and the MD analyses carried out to date indicate that these features tend to co-occur cross-linguistically associated with conversation and other (inter)personal spoken registers. Similarly, nouns, adjectives, and various kinds of nominal modifiers are found in many languages, and they tend to co-occur cross-linguistically associated with formal expository writing. By contrast, other linguistic resources are more specialised, occurring in comparatively few languages, and these resources have come to be exploited for more specialised and more distinctive dimensions of register variation.

The MD analysis of Spanish has illustrated the importance of both kinds of register patterns. To date, most comprehensive analyses of register variation have focused on English, making it difficult to determine which patterns are 'unmarked' (and candidates for cross-linguistic universals), and which patterns are more distinctive. It is only through analysis of a much wider range of languages, representing the full spectrum of typological and cultural differences, that we will be able to document fully the existence of cross-linguistic universals for register variation.

References

Alcalde, L. and R. Esperanza. 1999. 'Las intervenciones parlamentarias: lengua oral o lengua escrita?', Anuario de Estudios Filológicos 22, pp. 9-36.

Arce Castillo, A. 1999. 'Intensificadores en español coloquial', Anuario de Estudios Filológicos 22, pp. 37-48.

Ballester, A. and C. Santamaria. 1993. 'Transcription conventions used for the corpus of spoken contemporary Spanish', Literary and Linguistic Computing 8, pp. 283-92.

Besnier, N. 1988. 'The linguistic relationships of spoken and written Nukulaelae registers', Language 64, pp. 707-36.

Biber, D. 1985. 'Investigating macroscopic textual variation through multi-feature / multi-dimensional analyses', Linguistics 23, pp. 337-60.

Biber, D. 1986. 'Spoken and written textual dimensions in English: Resolving the contradictory findings', Language 62, pp. 384-414.

Biber, D. 1988. Variation across speech and writing. Cambridge: Cambridge University Press.

Biber, D. 1992. 'On the complexity of discourse complexity: A multidimensional Analysis', Discourse Processes 15, pp. 133-63.

Biber, D. 1995. Dimensions of register variation: A cross-linguistic comparison. Cambridge: Cambridge University Press.

Biber, D., Conrad, S. and R. Reppen. 1998. Corpus linguistics: Investigating language structure and use. Cambridge: Cambridge University Press.

Biber, D. and M. Hared. 1992. 'Dimensions of register variation in Somali', Language Variation and Change 4, pp. 41-75.

Biber, D. and M. Hared. 1992. 'Literacy in Somali: Linguistic consequences', Annual Review of Applied Linguistics 12, pp. 260-82.

Biber, D., and M. Hared. 1994. 'Linguistic correlates of the transition to literacy in Somali: Language adaptation in six press registers' in D. Biber and E. Finegan (eds.) Sociolinguistic perspectives on register, pp. 182-216. Oxford: Oxford University Press.

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

Bosque, I. and V. Demonte (eds.). 1999. Gramática descriptiva de la lengua española (3 volumes). Madrid: Real Academia Española

Brizuela, M., Andersen, E. and L. Stallings. 1999. 'Discourse markers as indicators of register', Hispania 82, pp. 128-41.

Butler, C.S. 1992. 'A corpus-based approach to relative clauses in the spoken Spanish of Madrid', Hispanic Linguistics 5, pp. 1-42.

Butler, C.S. 1998. 'Collocational frameworks in Spanish', International Journal of Corpus Linguistics 3, pp. 1-32.

Butt, J. and C. Benjamin. 2000 (third edition). A new reference grammar of modern Spanish. Chicago: NTC Publishing Group.

Collins, P. 1991. Cleft and pseudo-cleft constructions in English. London: Routledge.

Conrad, S. and D. Biber (eds). 2001. Variation in English: Multi-Dimensional studies. London: Longman.

Davies, M. 1995. 'Analyzing syntactic variation with computer-based corpora: The case of modern Spanish clitic climbing', Hispania 78, pp. 370-80.

Davies, M. 1997. 'A corpus-based analysis of Spanish subject raising in modern Spanish', Hispanic Linguistics 9, pp. 33-63.

Davies, M. 2002. Un corpus anotado de 100.000.000 palabras del español histórico y moderno. *SEPLN 2002 (Sociedad Española para el Procesamiento del Lenguaje Natural)*, 21-27.

Davies, M. 2003. 'Diachronic shifts and register variation with the "Lexical subject of infinitive" construction. (Para yo hacerlo)' in S. Montrul

and F. Ordóñez (eds.) Linguistic Theory and Language Development in Hispanic Languages, pp. 13-29. Somerville, MA: Cascadilla Press.

De Mello, G. 1992a. '*Le*' for '*les*' in the spoken educated Spanish of eleven cities', Canadian Journal of Linguistics, 37, pp. 407-30.

De Mello, G. 1992b. '*Se los*' for '*se lo*' in the spoken cultured Spanish of eleven cities', Hispanic Journal 13, pp. 165-79.

De Mello, G. 1995. 'El dequeísmo en el español hablado contemporáneo: ¿Un caso de independencia semántica?', Hispanic Linguistics 6-7, pp. 117-52.

De Mello, G. 2002. 'Leismo in contemporary Spanish American educated speech', Linguistics 40, pp. 261-83.

De Mello, G. 2004. 'Clitic doubling of postverbal direct objects: Lo tengo el anillo', Hispania 87, pp. 336-49.

Ferguson, C.A. 1983. 'Sports announcer talk: Syntactic aspects of register variation', Language in Society 12, pp. 153-72.

Gibbons, J. 1999. 'Register aspects of literacy in Spanish', Written Language and Literacy 2, pp. 63-88.

Hymes, D. 1984. 'Sociolinguistics: Stability and consolidation', International Journal of the Sociology of Language 45, pp. 39-45.

Jang, Shyue-Chian. 1998. Dimensions of spoken and written Taiwanese: A corpus-based register study. PhD dissertation. University of Hawaii.

Kaltenböck, G. 2005. '*It*-extraposition in English: A functional view', International Journal of Corpus Linguistics 10, pp. 119-60.

Kim, Y. 1990. Register variation in Korean: A corpus-based study. Unpublished doctoral dissertation, University of Southern California.

Kim, Y-J. and D. Biber. 1994. 'A corpus-based analysis of register variation in Korean' in D. Biber and E. Finegan (eds.) Sociolinguistic perspectives on register, pp. 157-81. Oxford: Oxford University Press.

Lope Blanch, J.M. (ed.). 1977. Estudios sobre el español hablado en las principales ciudades de América. México City: Universidad Nacional Autónoma.

Lope Blanch, J.M. 1991. Estudios sobre el español de México. Mexico City: Universidad Nacional Autónoma de México.

Marcos-Marín, F. 1994. Informática y Humanidades. Madrid: Gredos.

Ocampo, F. 1995. 'Pragmatic factors in word order: Constructions with a verb and an adverb in spoken Spanish', Probus 7, pp. 69-88.

Oh, S.-Y. 2000. '*Actually* and *in fact* in American English: A data-based analysis', English Language and Linguistics 4, pp. 243-68.

Parodi, G. 2005. 'Lingüística de corpus y análisis multidimensional: Exploración de la variación en el Corpus PUCV-2003' in G. Parodi (ed.) Discurso Especializado e Instituciones Formadoras, pp. 83-126. Valparaíso, Chile: Ediciones Universitarias de Valparaíso.

Prince, E.F. 1978. 'A Comparison of *Wh*-clefts and *It*-clefts in Discourse', Language 54, pp. 883-906.

Sáiz, M. 1999. A cross-linguistic corpus-based analysis of linguistic variation. Manchester: UMIST PhD Dissertation.

Sedano, M. 1994a. 'Evaluation of two hypotheses about the alternation between *aquí* and *acá* in a corpus of present-day Spanish', Language Variation and Change 6, pp. 223-37.

Sedano, M. 1994b. 'Presencia o ausencia de relativo: explicaciones funcionales', Thesaurus 49, pp. 491-518.

Sigley, R. 1997. 'The influence of formality and channel on relative pronoun choice in New Zealand English', English Language and Linguistics 1, pp. 207-32.

Thibault, A. 1987. 'Le Preterito et l'antepresente en espagnol dans la langue journalistique', Langues et Linguistique 13, pp. 287-320.

Tottie, G. 1991. Negation in English Speech and Writing: A Study in Variation. San Diego: Academic Press.

Ure, J. 1982. 'Introduction: Approaches to the study of register range', International Journal of the Sociology of Language 35, pp. 5-23.

Appendix A

Final factor structure with promax rotation

Linguistic variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
Indicat	0.86	-0.04	0.13	0.13	0.08	0.00
Existv	0.84	0.10	-0.14	-0.09	-0.14	0.02
causalsubord	0.83	-0.22	-0.07	-0.03	-0.07	0.03
Rbtime	0.76	-0.12	0.10	0.05	0.11	-0.02
Firstpro	0.76	-0.04	0.11	0.18	0.04	0.01
Mvser	0.74	0.02	-0.09	-0.10	-0.25	0.01
Demonpro	0.71	-0.03	-0.12	0.11	-0.09	0.03
othersingconj	0.70	-0.18	0.14	0.00	-0.15	0.02
prodrop1	0.69	0.24	0.04	0.14	0.12	0.00
mvestar	0.67	0.10	-0.05	0.12	0.08	-0.02
mentalv	0.65	0.12	0.10	0.20	0.01	-0.01
rbplace	0.63	-0.16	0.05	0.12	0.04	-0.02
exhaber	0.63	0.20	-0.21	-0.14	-0.04	-0.05
quevcompind	0.62	0.33	-0.12	-0.06	0.13	0.01
tagquest	0.61	-0.27	-0.18	0.20	-0.01	0.01
present	0.58	0.21	-0.27	0.20	-0.35	0.00
ira	0.54	0.20	-0.11	0.21	0.08	0.00
perfect	0.52	0.17	-0.07	-0.10	0.13	-0.07
communv	0.51	0.17	0.05	0.02	0.14	0.09
thirdpro	0.51	0.09	0.45	0.15	0.01	-0.01
progress	0.48	0.31	-0.16	-0.11	0.03	-0.01
oth_elque_all	0.47	0.21	-0.02	-0.14	-0.07	0.00
yesnoquest	0.47	0.22	-0.12	0.40	0.06	-0.07
querelindic	0.46	0.03	-0.01	-0.31	-0.05	0.03
rbmanner	0.45	-0.10	-0.01	0.24	-0.01	0.01
augment	0.41	-0.13	-0.01	-0.03	0.02	-0.01
quantifier	0.40	0.05	-0.03	-0.18	-0.20	-0.02
cuvbcomp	0.40	-0.04	0.09	0.17	-0.05	0.03
premoddem	0.38	0.24	-0.12	-0.23	-0.03	0.07
condexcepsubord	0.38	0.10	0.05	0.20	-0.19	0.02
tupro	0.37	-0.15	0.02	0.60	0.01	0.03
udpro	0.36	0.25	-0.03	0.00	0.05	-0.06
desirev	0.36	0.29	0.16	0.28	0.00	-0.01
gensingconj	0.35	-0.45	-0.04	0.10	-0.01	-0.05
facilv	0.34	0.33	0.14	0.04	-0.03	-0.03
occurv	0.30	0.01	0.10	0.04	0.05	0.01
imperfct	0.28	-0.27	0.59	-0.12	0.15	-0.04
prodroptu	0.28	-0.12	0.02	0.65	0.04	0.02
que_cleft	0.28	0.09	-0.05	-0.04	-0.06	-0.04
evaladj	0.25	0.07	-0.02	-0.06	-0.07	-0.01
predadj	0.24	0.10	0.04	-0.11	-0.34	-0.02

multiconj	0.22	0.19	0.03	-0.09	-0.20	-0.01
cuquest	0.22	-0.03	-0.01	0.48	0.01	-0.01
obligation_v	0.18	0.41	0.02	0.01	-0.20	0.01
othermente	0.18	0.17	-0.08	-0.35	-0.17	0.08
subjunct	0.16	0.66	0.03	0.20	-0.12	-0.02
diminut	0.14	-0.14	0.22	0.40	0.00	-0.03
sereflex	0.14	-0.02	0.36	0.08	-0.01	-0.05
quevcompsub	0.14	0.49	0.04	0.05	-0.03	-0.01
aspectv	0.10	-0.05	0.38	-0.02	0.05	0.01
querelsubjunc	0.10	0.56	-0.10	0.04	-0.08	0.00
prep_pro	0.06	0.02	0.20	0.30	0.05	0.06
ncompque	0.06	0.29	-0.01	-0.12	-0.04	0.02
preterit	0.05	-0.17	0.40	-0.04	0.64	0.01
conditnl	0.04	0.42	0.09	-0.04	-0.03	-0.01
exclamat	0.03	0.03	0.06	0.59	0.05	-0.01
seemocion	0.03	0.03	0.33	0.00	-0.01	-0.03
jcompque	0.02	0.24	0.00	-0.11	-0.14	-0.04
cualrel	-0.01	0.02	0.00	-0.02	0.02	0.95
propern	-0.02	0.01	-0.12	0.09	0.80	0.01
vplusinf	-0.03	0.46	0.30	-0.07	-0.30	0.03
othelcual	-0.03	0.00	-0.01	0.00	0.04	0.94
conditionals	-0.04	0.30	0.06	-0.08	0.06	0.00
future	-0.04	0.39	-0.11	0.11	0.13	0.04
serjque	-0.08	0.24	0.03	-0.13	-0.23	0.00
justse	-0.09	0.08	0.40	0.07	-0.13	0.05
possessives	-0.13	0.07	0.49	0.02	0.27	0.00
concesssubord	-0.15	-0.14	0.15	-0.11	-0.05	-0.01
novorainf	-0.18	0.38	0.33	0.01	-0.20	-0.01
serpsvpor	-0.21	-0.07	-0.07	0.01	0.24	0.06
clitic	-0.24	0.21	0.70	0.18	-0.25	0.02
serpsvagtlss	-0.26	-0.03	-0.05	-0.04	0.15	0.07
cuyo_rel	-0.27	-0.06	0.01	0.03	0.08	0.08
se_passive	-0.39	-0.26	-0.19	0.00	-0.21	0.00
other_adj	-0.45	-0.08	-0.20	-0.04	0.08	-0.01
avgwrdleng	-0.49	0.11	-0.23	-0.03	0.47	-0.01
premodadj	-0.51	-0.01	-0.02	-0.04	0.29	-0.02
postnompp	-0.54	-0.17	-0.11	0.09	-0.08	0.01
typtok	-0.57	0.13	0.27	-0.13	0.28	-0.03
derivedn	-0.58	0.24	-0.32	-0.12	0.01	0.02
nodetnp	-0.61	-0.22	-0.15	0.47	0.04	0.03
pluraln	-0.63	-0.16	-0.24	-0.03	-0.18	-0.06
preps	-0.64	0.01	-0.11	-0.02	0.42	0.02
defart	-0.67	-0.04	-0.16	-0.03	0.20	-0.04
postmodadj	-0.69	-0.06	-0.29	-0.01	0.07	-0.03
singnoun	-0.76	-0.05	0.07	0.05	-0.06	0.00

	Eigenvalues for the first six factors			
	Eigenvalue	*Difference*	*Proportion*	*Cumulative*
1	23.5405150	18.3460891	0.2737	0.2737
2	5.1944259	1.8734604	0.0604	0.3341
3	3.3209655	0.7451271	0.0386	0.3727
4	2.5758384	0.2986935	0.0300	0.4027
5	2.2771448	0.3051511	0.0265	0.4292
6	1.9719937	0.1077713	0.0229	0.4521

	Inter-Factor Correlations					
	Fact 1	*Fact 2*	*Fact 3*	*Fact 4*	*Fact 5*	*Fact 6*
Factor 1	1.00	0.26	0.27	0.44	-0.36	-0.14
Factor 2	0.26	1.00	-0.03	-0.02	-0.15	-0.06
Factor 3	0.27	-0.03	1.00	0.19	-0.05	-0.08
Factor 4	0.44	-0.02	0.19	1.00	-0.24	-0.10
Factor 5	-0.36	-0.15	-0.05	-0.24	1.00	0.02
Factor 6	-0.14	-0.06	-0.08	-0.10	0.02	1.00

Copyright of Corpora is the property of Edinburgh University Press and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.