A multi-dimensional analysis of register variation in Brazilian Portuguese

Tony Berber Sardinha,[1] Carlos Kauffmann[2] and Cristina Mayer Acunzo[3]

Abstract

In this paper, we present a Multi-Dimensional analysis of Brazilian Portuguese, based on a large, diverse corpus comprising forty-eight different spoken and written registers. Previous research in MD analysis includes multi-register investigations of a range of languages, including English, Spanish, Somali and Korean, among others. At the same time, a large body of literature on text varieties in Brazilian Portuguese exists, but previous research focusses on specific aspects of one, or at the most, a few varieties at a time and, therefore, does not present a comprehensive picture of register use in the linguistic community of Brazilian Portuguese speakers. In this study, we attempt to fill this gap by employing the MD framework, enabling researchers to account for a large number of different registers, based on a wide repertory of linguistic features. The analysis revealed six dimensions of variation, which are presented, illustrated and discussed here.

Keywords: Brazilian Portuguese, dimensions of register variation, Multi-dimensional analysis, register analysis

1. Introduction

Multi-Dimensional (MD) analysis is a corpus-based method introduced by Douglas Biber (1985, 1986, 1988) that uses multi-variate statistical techniques to identify the salient co-ocurrence patterns of linguistic characteristics across registers. These patterns, called dimensions, represent

[1] LAEL, PUCSP, São Paulo Catholic University, R. Monte Alegre 984, São Paulo, SP 05014-001, Brazil.

[2] R. Camburiú 52, São Paulo, SP 05058-020, Brazil.

[3] R. Padre Raposo 881 apt. 31, São Paulo, SP 03118-001, Brazil.

Correspondence to: Tony Berber Sardinha, *e-mail*: tonycorpuslg@gmail.com

Corpora 2014 Vol. 9 (2): 239–271
DOI: 10.3366/cor.2014.0059
© Edinburgh University Press
www.euppublishing.com/cor

the underlying parameters of register variation for the text varieties that are being investigated. The notion that texts can be grouped together based on shared patterns of linguistic features pre-dates corpus linguistics, going back at least to Carroll (1960), who identified stylistic 'vectors' in literature (Biber, 2014). Inspired by this idea, Biber formulated the MD method for approaching cross-register variation from a corpus perspective. Since its inception, the MD framework has been used to analyse a growing number of registers from different contexts, such as academia (Atkinson, 2001; and Shergue, 2003), the press (Condi de Souza, 2014; and Kauffmann, 2005), university (Biber, 2006), language teaching and learning (Conde, 2002; Crossley *et al*., 2014; Delegá-Lúcio, 2013; and Reppen, 2001), cinema (Veirano Pinto, 2014), television (Quaglio, 2009), music (Bértoli Dutra, 2014), aviation (Zuppardo, 2014), engineering (Conrad, 2014) and electronic communication (Berber Sardinha, 2014; Biber and Conrad, 2009; Biber and Kurjian, 2007; Grieve *et al*., 2010; and Titak and Roberson, 2013), among others. It has also been used to account for entire languages, including English (Biber, 1988; Crossley and Louwerse, 2007; de Mönnink *et al*., 2003; and Lee, 1999), Korean (Kim and Biber, 1994), Somali (Biber and Hared, 1994), Nukulaelae Tuvaluan (Besnier, 1988), Gaelic (Lamb, 2008) and Spanish (Biber *et al*., 2006; Biber and Tracy-Ventura, 2007; and Parodi, 2007).

As with most modern corpus linguistic approaches, the primary goal of the MD analysis is to investigate language in use and, in so doing, to deliver rich, detailed, data-based descriptions of the multi-faceted ways in which human beings use language in society. However, unlike other approaches, the MD analysis has a firm belief in the variation inherent in language use; its theoretical credentials and method pay tribute to the founding principle that language is both extravagant and contextually patterned; and, thus, the job of the MD analyst becomes one of making sense of the variation. It also recognises that language use is profoundly complex, emerging from multiple interactions among numerous linguistic features and situational characteristics; therefore, this complexity can only be modeled by a multi-dimensional space where all these multiple influences operate at once to shape language events.

Conceptually, the MD research framework is part of a larger research programme developed by Biber over the years to describe association patterns, or 'the systematic ways in which linguistic features are used in association with other linguistic and non-linguistic features' (Biber, 2000: 289). Association patterns can be of three kinds: for individual words, for grammatical features and for analysing register variation. MD analysis reflects the last of these three facets and serves to investigate 'the variability among texts' through dimensions or the 'co-occurrence patterns of linguistic features' (Biber, 2000: 289). Although the first two of these are devoted to 'investigating the variability of a linguistic feature in terms of its association patterns' through 'linguistic associations' (i.e., lexis or grammar), MD analysis is concerned with 'non-linguistic association patterns', which 'describe how certain linguistic features are differentially associated with registers, dialects, or historical change' (Biber, 2000: 290).

Texts and registers are two major cornerstones of the MD analysis. The text is the central unit of analysis in MD research; thus, all linguistic observations are recorded with respect to their occurrence in individual texts. In other corpus linguistic approaches, the boundaries between texts in the corpus can be erased and the corpus treated as a 'homogenous blob' (Biber, 2013a: 371); by contrast, the divisions between texts are maintained throughout in MD analyses. One of the simplest manifestations of the allegiance to the text is the fact that examples in MD analyses are normally given as extended text samples and rarely, if ever, as concordance lines, because in these the boundaries between texts are, by default, lost. The MD analysis is one of the exponents of the text-linguistic approach to the study of linguistic variation and use (Biber, 2012: 33), where the 'research goal is to describe the characteristics of texts, rather than the characteristics of a linguistic feature' (Biber, 2012: 33).

Registers, in turn, are situationally defined text categorisations, which can be of any level of specificity – from very broad classes, like official documents and academic prose (Biber, 1988), down to very fine-grained categories, like civil engineering project reports (Conrad, 2014) or commercial aviation maintenance manuals (Zuppardo, 2014). In an MD corpus design, texts are chosen in order to, and because they, represent particular registers; in the ensuing MD analysis, texts are analysed as exemplars of the register. Register has been proposed as a central construct in corpus linguistic research (Biber, 2012) and as the driving force behind the analysis, rather than as an afterthought: 'the practice advocated [. . .] is to begin a research study with the hypothesis that [. . .] register differences exist, and to include analysis of those differences unless they are empirically shown to be unimportant' (Biber, 2012: 34). As a result, the register comparisons, one of the hallmarks of MD research, directly reflect the central role played by registers in shedding light on the variation that is inherent in language use.

In this paper, we follow this tradition and report on a language-wide MD analysis of Brazilian Portuguese – the most populous variant of Portuguese. Portuguese is a Romance language with more than 200 million users around the world and ranks sixth in terms of its number of native speakers, of which approximately 90 percent are Brazilian.[4] Several distinctive differences exist among the varieties of Portuguese in terms of lexis (Bacelar do Nascimento *et al.*, 2014; and Kilgariff *et al.*, 2014), syntax (Castilho, 2009: 883–5) and register use, which led us to restrict the scope of our study of this language to a single national variety.

Numerous studies have examined the textual varieties of Brazilian Portuguese from a range of perspectives, including discourse analysis (Souza e Silva and Brait, 2013), variation (Preti, 2005) and corpus-based

[4] See: www.ethnologue.com

(Berber Sardinha, 2005). However, most studies on the registers and genres of Brazilian Portuguese tend to analyse one variety, or a few varieties, at a time, generally focussing on a small set of linguistic characteristics. The main goal of this study is to fill this gap by looking at a wide range of registers, both written and spoken, and based on a broad set of linguistic features. In so doing, we will identify the dimensions underlying register variations in Brazilian Portuguese. The only other MD analysis of Brazilian Portuguese to date is Kauffmann's (2005) analysis of newspaper registers, based on a carefully constructed corpus comprising fourteen different registers, which identified two dimensions: Narrative *versus* Expository, and Argumentation *versus* Informational.

2. Methods

Following the standard method proposed by Biber (1988), we revised the existing literature on structural characteristics of Portuguese (e.g., Azevedo, 2005; Bechara, 1999; Castilho, 1989; Cunha, 2001; Ilari, 1991; Moura Neves, 2000; Thomas, 1969; and Whitlam, 2010), collected a representative corpus, tagged it electronically for part-of-speech and lexical features, collected the frequencies of the individual features, processed these counts statistically, and interpreted the resulting factors qualitatively in terms of the discourse functions performed by the co-occurring groups of features. Our corpus, named the Brazilian Register Variation Corpus (CBVR; Corpus Brasileiro de Variação de Registro), was compiled specifically for this research project, but it has since been used in other projects (e.g., Berber Sardinha *et al*., 2014). It contains forty-eight different registers (5,644,006 words total): twelve spoken (1,547,853 words / 27.5 percent) and thirty-six written (4,096,153 words / 72.5 percent), being the largest in terms of text varieties when compared with other corpora used in MD analyses (e.g., English, Tuvaluan, Korean, Somali and Spanish). The composition of the corpus can be seen under Appendix A. The written component was collected from both print and online sources, with additional material typed up, and the spoken component was drawn from both online sources and existing collections of transcribed speech (Projects Iboruna, Porcufor and Museu da Pessoa). All texts were cleaned up using scripts that were especially designed for this project, then edited and checked by hand so as to render them as tagger-friendly as possible by normalising punctuation, fixing spelling errors and capitalisation inconsistencies, and keeping typographical features to a minimum. The corpus was tagged using PALAVRAS, a state-of-the-art parser for Portuguese that automatically annotates more than 300 different features, including morphology, syntax and semantics, with an accuracy rate of 98.6 percent for parts of speech, 95 percent for syntax and 99 percent for lemmatisation (Bick, 2014: 298). The final inventory comprises 190 features, of which ninety-three remained after the final factor extraction. Individual feature counts were normed per 1,000 words to enable comparisons of texts of different sizes.

An initial factor analysis was run on the normed data, using principal axis factoring as the extraction method. A scree plot was generated and, upon inspection, it suggested the existence of six factors in the data (see Figure 1). These six factors account for 31 percent of the variation in the corpus (see Table 1 for the amount of variance captured by each factor). A rotated factor

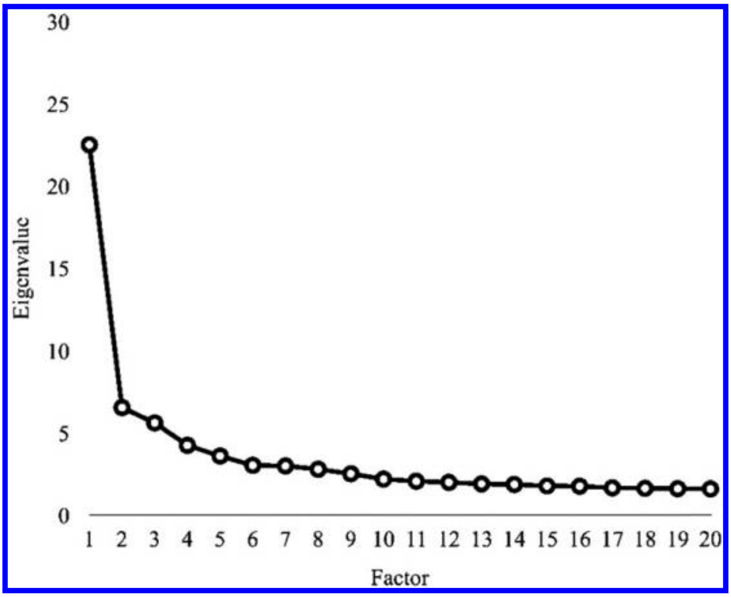

Figure 1: Scree plot

Factor	Eigenvalue	Percent of variance	Cumulative percent
1	22.514	15.316	15.316
2	6.554	4.458	19.774
3	5.619	3.823	23.597
4	4.257	2.896	26.493
5	3.606	2.453	28.946
6	3.039	2.067	31.014
7	3.003	2.043	33.056
8	2.8	1.905	34.961
9	2.527	1.719	36.68
10	2.197	1.495	38.175

Table 1: Eigenvalues and variance accounted for by the first ten factors

Factor	1	2	3	4	5	6
1	1.000	0.165	0.482	0.235	−0.412	0.156
2	0.165	1.000	0.045	−0.027	−0.119	−0.148
3	0.482	0.045	1.000	−0.112	−0.362	0.232
4	0.235	−0.027	−0.112	1.000	−0.056	−0.193
5	−0.412	−0.119	−0.362	−0.056	1.000	−0.200
6	0.156	−0.148	0.232	−0.193	−0.200	1.000

Table 2: Inter-factor correlations

analysis was subsequently performed, with Promax rotation, which allows for some correlation among the factors. The factor inter-correlations (shown in Table 2) are weak, ranging from −0.41 to 0.48. These values fall within the normal range for MD studies. A cut-off was applied to the factor weights, whereby only those features with loadings equal to or greater than +/− 0.3 were used (following Biber, 2006: 14). A feature was only entered once in the computation of the dimension score – namely, for the factor on which it had the greatest absolute weight. In Tables 3 to 8, features that have larger weights on a different factor are enclosed in parentheses.

Each factor was then interpreted functionally, based on the shared social, communicative and discourse characteristics of the features on it (see Biber, 1995: 136–8). The interpretation of factors leads to the definition of the dimensions and is, therefore, a crucial step in an MD analysis. In addition to considering the functions of linguistic features in the interpretation of the factors, the analysts should also ‘consider the similarities and differences among registers with respect to the set of co-occurring linguistic features’ (Biber *et al.*, 2006: 14). This requires the computation of dimension (or factor) scores for each text, by summing up the standardised scores of the features loading on each pole of a dimension and then subtracting the sum of the negative pole features from the positive ones. (If a dimension has a single pole, the subtraction is not necessary.)

3. Interpretation of the factors

Factor 1 (Table 3) incorporates the greatest number of features, with forty-nine linguistic characteristics in total (thirty-five positive and fourteen negative), which accounts for the greatest variation (15.3 percent). The positive features are mostly verbs (ten features), adverbs (seven) and pronouns (seven). First-person verb forms and first-person pronouns (either in subject or object positions) both place a focus on the addressor, whereas second-person pronouns indicate an addressee focus. Both *QU* and yes–no questions generally indicate turn initiation between addressor and addressee

Feature [label]	Loading
Pronouns: Second person, in object position [prn2obl]	0.910
Verbs: First person [vb1]	0.826
Verbs: Mental [vbment]	0.827
Verbs: *Ir* future [vbfutir]	0.821
Pronouns: Second person singular, in subject position [prn2sngsubj]	0.763
Verbs: Private [vbpriv]	0.702
Verbs: Action [vbact]	0.689
Adverbs: *Não* (no) [advnao]	0.674
Pronouns: First person singular, in subject position [prn1sngsubj]	0.633
Adverbs: Time [advtime]	0.629
Pronouns: First person, object position [prn1obl]	0.622
Pronouns: Quantifier [prnqtf]	0.616
Adjectives: Evaluative [adjeval]	0.612
Pronouns: Possessive [prnposs]	0.605
Adverbs: Intensity [advints]	0.552
QU questions [qsqu]	0.539
Adverbs: Amplifier [advampl]	0.522
Adverbs: Emphatic [advemph]	0.476
Que clause controlled by verb in the indicative [vbqueindic]	0.467
(Adverbs: Place [advpl]	0.460)
Adjectives: Predicative position [adjpred]	0.456
(Yes or no question [qsyn]	0.421)
Verbs: Infinitive [vbinf]	0.407
Adverbs: Manner [advmanner]	0.373
Verbs: Communication [vbcomm]	0.356
(Discourse marker [discmrkr]	0.350)
Verbs: Gerund form, all [vbgerall]	0.345
(Subject omission [subjdrop]	0.343)
Modals: *Precisar* (need to) [mdprecisar]	0.338
Que clause controlled by adverb [clqueeadv]	0.337
Verbs: Progressive preceded by infinitive [vbproginf]	0.330
(Verbs: Future subjunctive [vbsubfut]	0.323)
Adverbs: Negative, except *não* [advneg]	0.320

Table 3: Loadings on Dimension 1: Oral *versus* Literate Discourse. (*Note*: Features with larger weights on a different factor are enclosed in parentheses)

Feature [label]	Loading
(Subordinating (conditional) clause [cjcond]	0.313)
Pronouns: Nominal in subject position [prnnomsubj]	0.308
Reduced progressive clause [vbprogphr]	−0.309
Pronouns: Relative *qual* or *cujo* [prnqualcujo]	−0.315
(Adjectives: Affiliative [adjaffi]	−0.336)
Agentless passives [clpassless]	−0.377
Adjectives: Relational [adjrela]	−0.422
Past participle [vbpastprt]	−0.497
Nominalisation in subject position [nominlzsubj]	−0.505
Adjectives: Topical [adjtopi]	−0.511
Nouns: Abstract [nabst]	−0.521
Average word length [wl]	−0.529
Adjectives: Attributive position [adjattr]	−0.589
Nouns: Compound [ncomp]	−0.651
Articles: Definite [artdef]	−0.739
Prepositions: All [prpall]	−0.776

Table 3: (*Continued.*)

and are used ‘primarily in interactive discourse’ (Biber, 1988: 106). Mental and private verbs are used to express one’s inner thoughts and feelings, thereby contributing to the ‘addressee focus’, while communication verbs shift that focus to the other participants in the discourse. Possessives can also contribute to the personal nature of the dimension, as they signal how entities relate to individuals in terms of ‘ownership’. Action verbs, on the other hand, highlight the events taking place in a more ‘concrete’ way than the mental and private verbs, which can perform a more ‘subjective’ role; however, these are frequently voiced in the first person, thereby reinforcing the idea of an interplay between addressor and addressee. Progressive forms preceded by an infinitive (e.g., *vou estar fazendo*, ‘I will be doing’) mark an informal vernacular, sometimes uneducated, future form. A related kind of informal future is the *ir* future (e.g., *vou fazer*, ‘I’m going to do’), which is typically associated with spontaneous discourse. Negation adverbs (both simple *não* and other related forms such as *nunca* and *jamais*) convey denials and rejections, which can signal how interlocutors negate propositions. Intensity adverbs, amplifiers and emphatics all ‘mark heightened feeling’ (Biber, 1988: 106) and again contribute to shaping the discourse as person-orientated. Place, manner and time adverbs are all ‘other oriented features’, ‘reflecting the description of other people in particular places and times’ (Biber *et al.*, 2006: 12), thereby fitting this understanding of the dimension as combining both a 'self-' and an 'other-'orientated perspective.

The negative pole of factor 1 has mostly nominal features, such as nouns, adjectives, articles and prepositions, which, in general, contribute to presenting information in a condensed manner in the text. Nouns are the 'primary bearers of referential meaning in a text' (Biber, 1988: 104), and a concentration of them increases informational density. In nominal groups, prepositions connect the individual elements and help integrate information in compact units. Past participials are used in agentless passive-voice constructions, which contribute to raising the level of abstractness in the discourse by eliding the agent. Reduced progressive clauses also work in a similar fashion: simultaneously, they remove agency and increasingly condense information by latching extra information onto the main clause. Abstract nouns and nominalisation boost the abstract nature of the information. Adjectives qualify the information being presented and, in the attributive position, they further increase the amount of information in the nominal group. The relative pronouns *qual* and *cujo* are both markers of learned discourse, which is typically characterised by high levels of abstraction and information content. Long words mark precise lexical choices, such as those instantiated by abstract nouns, nominalisations, participles and gerund forms. Based on the functional associations of the features on the register distinctions, we posit the interpretive label 'Oral *versus* Literate Discourse' for Dimension 1, with orality representing 'on-line production, and the literate pole [...] careful, usually edited, written production' (Biber, 1995: 242). The mean dimension scores are presented in Figure 2.

Although the positive end of Dimension 1 includes both written and spoken varieties, all but two of the spoken registers are featured there. Most highest-scoring registers are addressee-orientated, dialogue-based texts, such as face-to-face conversation, soap opera scripts and interviews, but among these are also 'simulated conversation' registers such as songs and Facebook messages. Example 1 illustrates the dense use of involved features, such as first-person verb forms (*fiz*, etc.), first- (*eu*) and second-person (*te*) pronouns, mental (*acreditei*) and action verbs (*foi*, etc.), and *ir* future forms in songs (*ia passar*).

(*1*) *Eu* já *fiz* de tudo pra *te* convencer
Mandei rosas vermelhas
Lindas pra *você*
Falei de amor
Fiz uma canção
A Lua se *foi*, nem *vi* o sol *chegar*
Acreditei que o tempo não *ia passar*
[I've already done everything to convince you
Sent you red roses

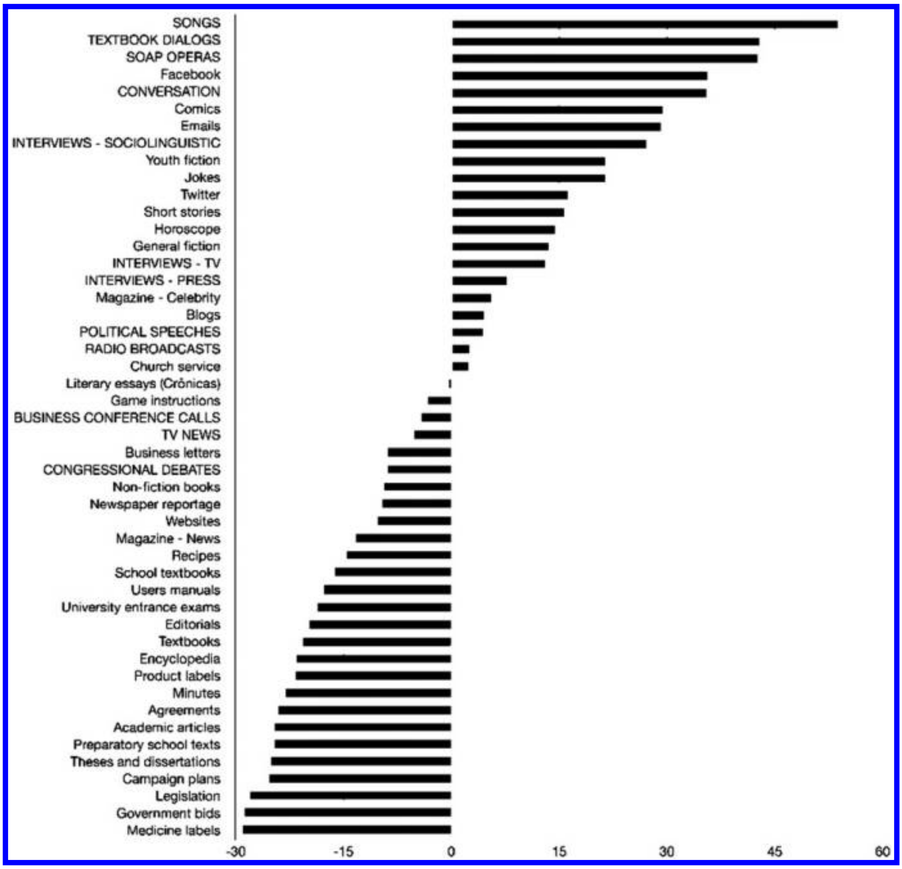

Figure 2: Dimension 1: Oral *versus* Literate Discourse

Beautiful ones for you
Spoke about love
Wrote a song
The moon is gone, I haven’t seen the sun arrive]

On the opposite extreme are mostly written registers, loaded with informational content conveyed through prepositions (*para*, etc.), abstract nouns (*indicações*) and noun compounds (*composição de doses*, etc.), nominalisations (*tratamento*), adjectives in attributive position (*sistêmico*), and past participles (*indicado*), among other features. Example 2 is a sample from drug labels that illustrate the use of these features.

(*2*) *Os* comprimidos apresentam dois sulcos *para* facilitar *a composição de doses*. *Indicações*: [medicine name] é um *glicocorticóide de uso oral*, *indicado para tratamento de diversas patologias*, tais *como*: *Reumatologia*: artrite *reumatóide*, lupus *eritematoso sistêmico* (...)

Feature [label]	Loading
Que clause controlled by noun [nounque]	0.593
Pronouns: Relative *que* [prnque]	0.529
Adverbs: Comparative [advcomp]	0.473
Nouns: Cognition [ncogn]	0.451
Que or infinitive clause controlled by noun (stance) [nqueinfcl]	0.447
Infinitive clause controlled by adjective [clinfadj]	0.447
Que clause controlled by preposition [clqueeprp]	0.426
Pronouns: Demonstrative [prndem]	0.406
Infinitive clause controlled by preposition [clinfprp]	0.395
Que clause controlled by adjective (stance) [adjque]	0.378
(Adjectives: Predicative position [adjpred]	0.353)
(Pronouns: Quantifier [prnqtf]	0.352)
(Pronouns: Third person, object position [prn3obl]	0.343)
(Modals: *Poder* [mdpoder]	0.336)
Infinitive clause controlled by ease or difficulty adjective [clinfadjease]	0.334
Adverbs: Hedge [advhedg]	0.331
Articles: Indefinite [artindef]	0.325
Verbs: Future preterit tense [vbfutpret]	0.311
Conjunctions: Co-ordinating (adversative) [cjadv]	0.310

Table 4: Loadings on Dimension 2: Argumentation. (*Note*: Features with larger weights on a different factor are enclosed in parentheses)

> [The pills have two grooves that facilitate dosage adjustment. Prescription: [medicine name] is a glucocorticoid that is to be taken by mouth, prescribed for the treatment of numerous pathologies, such as: Rheumatology: rheumatoid arthritis, systemic lupus erythematosus (. . .)]

Factor 2 (see Table 4) comprises only positive features. Most of the nineteen linguistic characteristics are non-finite clauses, such as *que* (that/which) or infinitive clauses controlled by specific kinds of nouns or adjectives, or by prepositions. In addition, the factor includes relative/adjectival clauses introduced by *que* relative pronouns. Clauses controlled by stance nouns or adjectives or by a particular class of adjectives (expressing ease or difficulty) are regularly employed to ascertain a point of view or to frame information in a particular way. Relative clauses are ‘devices for the explicit, elaborated identification of referents in a text’ (Biber, 1988: 110), which can then be commented on in the text in a manner that suits one's framing. Demonstrative pronouns can also be used to single out particular referents, whether concrete or abstract. Comparative adverbs are commonly used as devices for evaluating entities or propositions and, as such, are valuable elements for expressing personal attitude. Similarly, quantifier pronouns enable comparisons and thereby function as rhetorical devices as well. Co-ordinating adversative conjunctions introduce clauses that also mark comparisons and contrasts. Cognition nouns are rhetorical devices that convey abstract notions; they enable speakers to encapsulate complex information into a single word, which in turn affords greater control over how the information will be handled to achieve one's intentions in the discourse. Hedges express fuzziness or vagueness and, as such, they can also function as rhetorical devices. Indefinite articles indicate unspecified entities, thereby denoting a degree of abstractness. The modal *poder* expresses a variety of meanings, which include (but are not limited to) ability (*posso levantar isso*, 'I can lift that'), likelihood (*pode chover*, 'it might rain') and permission (*Posso sair?* 'May I leave?'), all of which are useful for shaping propositions. Third-person pronouns in the object position mark a form of 'other direct discourse' detached from the immediate interlocutors. Adjectives in the predicative position are used to qualify particular entities. Future preterite forms are regularly used to formulate hypothetical statements. Put together, these features seem to mark the argumentative use of language; therefore, the proposed interpretive label for Dimension 2 is argumentation.

Figure 3 shows the register differences associated with Dimension 2. High scoring registers include not only standard argumentative debate-based registers such as political speeches, interviews (both print and TV), editorials and congressional sessions, but also horoscopes, which are not commonly regarded as argumentative. Low-scoring registers, on the other hand, are defined by an absence of such features and, hence, of 'Argumentation'.

The sample in Example 3 shows the use of features associated with argumentation in a horoscope, such as demonstrative pronouns[5] (*este*, etc.), relative clauses (*trilha que*…, etc.), adversative conjunctions (*mas*), the modal *poder*, and cognition nouns (*sentimentos*).

(*3*) Siga a *trilha que* seus *sentimentos* propuserem hoje. Permita-se a surpresa, deixe de lado todos os planejamentos *que* tiver feito para *este* dia e observe a *direção que* seus *sentimentos mais* íntimos estabelecerem. Conversas sérias *podem* e devem ser desenvolvidas, *mas* antes de tudo você deve tirar de cima *dessas* o ar de gravidade, pois sem *isso* só encontrará resistência (…)

[Follow the path that your feelings have laid out today. Allow yourself to be surprised, leave to one side all the planning that you

[5] In Brazilian Portuguese grammars, these are usually referred to as demonstrative pronouns, not as demonstrative determiners (e.g., Bechara, 1999).

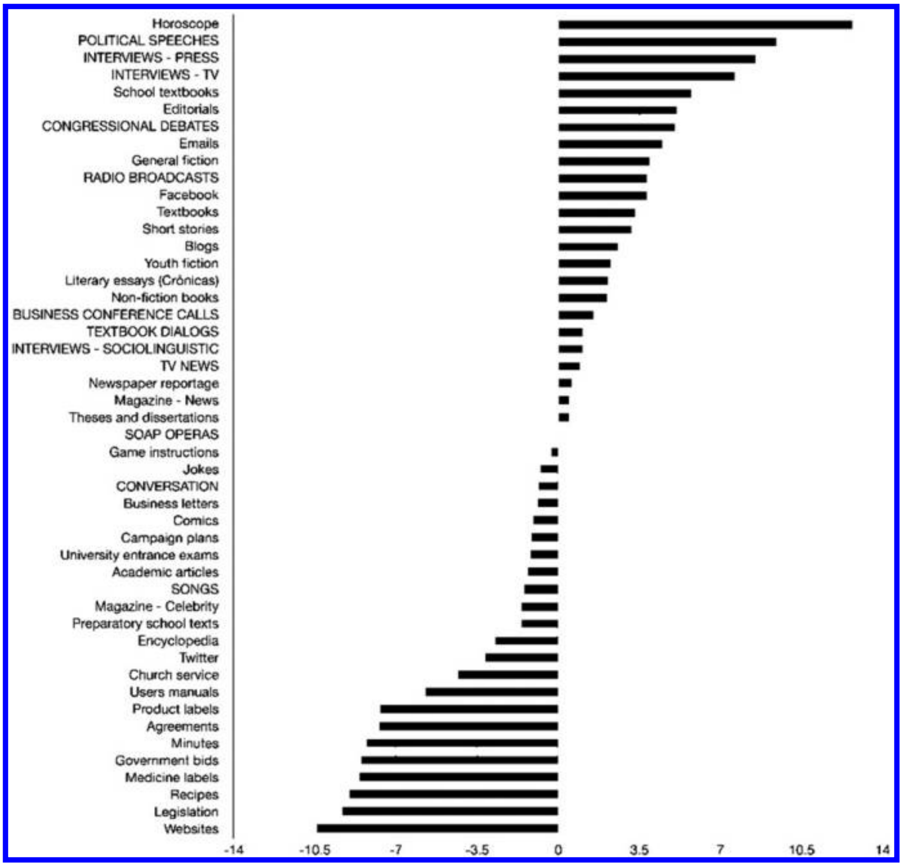

Figure 3: Dimension 2: Argumentation

> have done for this day and note the direction that your most intimate feelings established. Serious conversations can and should be carried out, but before anything else you should take that air of gravity off them, because without that you will only find resistance (. . .)]

Factor 3 (see Table 5) includes fifteen features on two poles, most of which (twelve) have positive loadings. Tag questions and yes–no questions are both addressee-orientated features. Tag questions are interactive devices that can invite feedback from an interlocutor without a direct question. Yes–no questions, on the other hand, explicitly signal a change of turn, directly requesting the participation of the addressee. Third-person personal pronouns in both the subject and object positions are ‘other-orientated’ features that refer to participants outside the immediate context of the interaction. Conclusive conjunctions mark a specific logical relationship between clauses, indicating an inferential or deductive form of reasoning. Place adverbs typically mark a deictic reference so that the discourse

Feature [label]	Loading
Tag questions [qsttag]	0.795
Contractions [contrac]	0.714
Discourse marker [discmrkr]	0.671
Questions: Yes or No question [qsyn]	0.547
Pronouns: Third person singular, in subject position [prn3sngsubj]	0.498
Pronouns: Third person plural, in subject position [prn3plusubj]	0.481
Conjunctions: Co-ordinating (conclusive) [cjcncl]	0.475
Adverbs: Place [advpl]	0.462
Modals: *Ter que / ter d*e (have to, ought to) [mdter]	0.315
(Pronouns: Demonstrative [prndem]	0.362)
(*Que* clause controlled by verb in the indicative [vbqueindic]	0.316)
(Pronouns: First person singular, in subject position [prn1sngsubj]	0.314)
Type–token ratio [ttr]	−0.346
(Adjectives: Attributive position [adjattr]	−0.357)
(Pronouns: Possessive [prnposs]	−0.431)

Table 5: Loadings on Dimension 3: Involved *versus* Informational Production. (*Note*: Features with larger weights on a different factor are enclosed in parentheses)

is grounded in the immediate context. *Ter que* and *ter de* are necessity modals, expressing the speaker's or writer's stance in terms of an obligation. Demonstrative pronouns signal a reference that is in the vicinity of the addressor. *Que* clauses controlled by a verb in the indicative mood are regularly reporting clauses or complements to a copula, such as *parecer* ('to seem/appear'). First-person personal pronouns refer directly to the speaker or writer. This set of features seems to point to a person-orientated, involved, interactive discourse that in many ways resembles the positive pole of both English (Biber, 1988) and Spanish Dimensions 1 (Biber *et al*., 2006; and Biber and Tracy-Ventura, 2007).

The negative pole of factor 3 has three features, only one of which has a higher loading on this factor – namely, the type–token ratio, which measures the variety of vocabulary that is used. Texts with high ratios have a wide range of lexical choices. Adjectives in the attributive position are 'used to further elaborate nominal information' (Biber, 1988: 105), reflecting a higher density of information. These features enable the conveyance of concentrated information. We therefore propose the interpretive label 'Involved *versus* Informational Production' to reflect the shared communicative functions of Dimension 3.

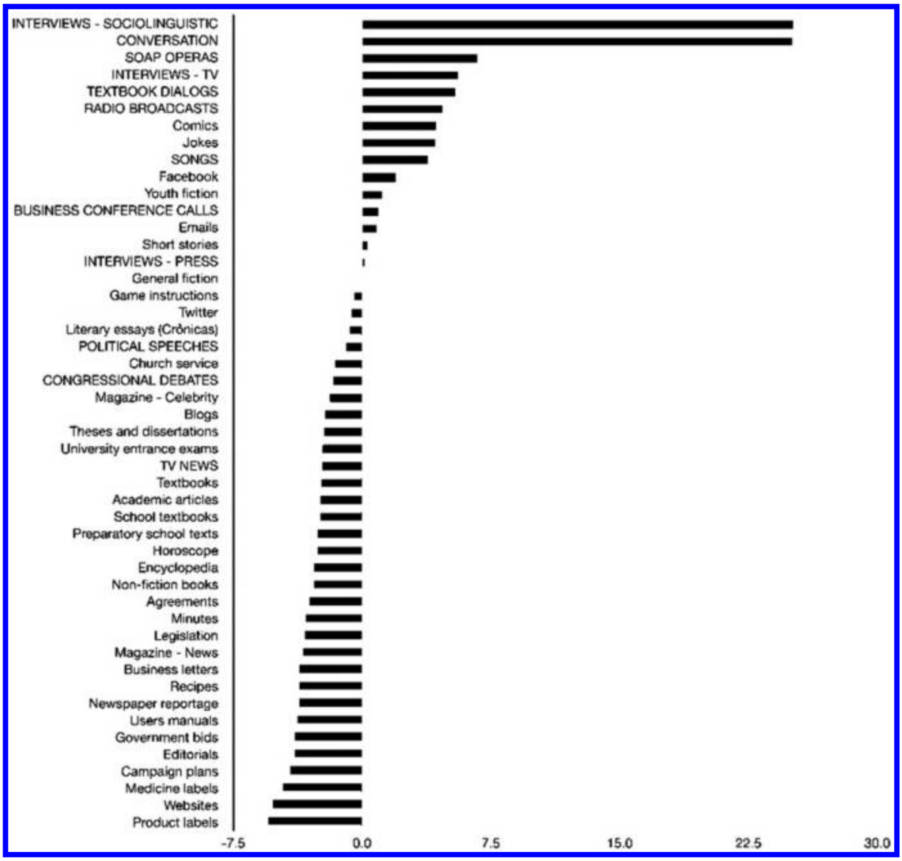

Figure 4: Dimension 3: Involved *versus* Informational Production

The mean registers scores for Dimension 3 appear in Figure 4. Sociolinguistic interviews and conversations are the most marked, followed by soap operas and TV interviews, thereby supporting the interpretation of the positive pole as reflecting involvement. At the same time, the registers on the negative pole all display high informational content (e.g., product labels, drug information leaflets (*bulas*) and legislation).

Example 4 is an excerpt of a sociolinguistic interview, in which question tags (*né*), place adverbials (*lá*) and discourse markers (*então*, etc.) are used.

(*4*) A: *Então*, o que que chocou na morte de Ayrton Senna, *né*? B: *É*, de repente, corrida, corrida pra mim seria o Ayrton Senna, *né*? *Então*, ligava-se a televisão pra assistir não a corrida propriamente dita, mas, de repente, pra ver alguém da gente, *né*? Brasileiro empunhando a bandeira *lá*.

[A: So, that was shocking, Ayrton Senna's death, wasn't it?
B: Yeah, I mean, a race, a race to me would have to be with Ayrton

Feature [label]	Loading
Verbs: Present subjunctive [vbsubpres]	0.821
Verbs: Imperative [vbimp]	0.774
Nouns: Concrete [nconc]	0.565
Subject omission [subjdrop]	0.545
Verbs: Facilitation [vbfacil]	0.485
Conjunctions: Co-ordinating (clausal) [cjcoorcls]	0.465
(Adverbs: Manner [advmanner]	0.345)

Table 6: Loadings on Dimension 4: Directive Discourse. (*Note*: Features with larger weights on a different factor are enclosed in parentheses)

> Senna, you know what I mean? So, you would turn on the TV not to watch the actual race, but, really, to see people like us, you know? Brazilians hoisting the flag over there.]

Factor 4 (Table 6) has seven features on a single pole. Six variables load primarily on it, including the present subjunctive and the imperative moods, both of which commonly occur in pro-drop structures to express directives. Concrete nouns refer to ‘inanimate objects that can be touched’ (Biber, 2006: 248), whereas facilitation or causation verbs ‘indicate that some person or inanimate entity brings about a new state of affairs’ (Biber, 2006: 247). Clausal co-ordination links clauses but retains their independent equal status. Adverbs of manner convey ‘information about how an action is performed’ (Biber *et al*., 1999: 553). The co-occurrence of these characteristics is frequently used to give instructions on how to perform tasks, many of which are ‘hands on’ and achieve particular outcomes, usually of a practical nature; hence, we propose the interpretive label ‘Directive Discourse’ for Dimension 4.

The register differences defined by Dimension 4 are shown in Figure 5; the most marked register is recipes, which by their very nature require directives to explain how to prepare drinks and dishes. Owner’s manuals and game instructions also provide directions for operating tools and appliances as well as for playing games.

Example 5 illustrates the use of imperatives (*lave*, etc.), concrete nouns (*limão*, etc.), and adverbs of manner (*bem*) in a recipe.

(5) *Caipirinha*
1 *limão-galego*
1 *colher* (*sopa*) *de açúcar*
2 *cubos de gelo* triturados
4 *colheres* (*sopa*) de *pinga*

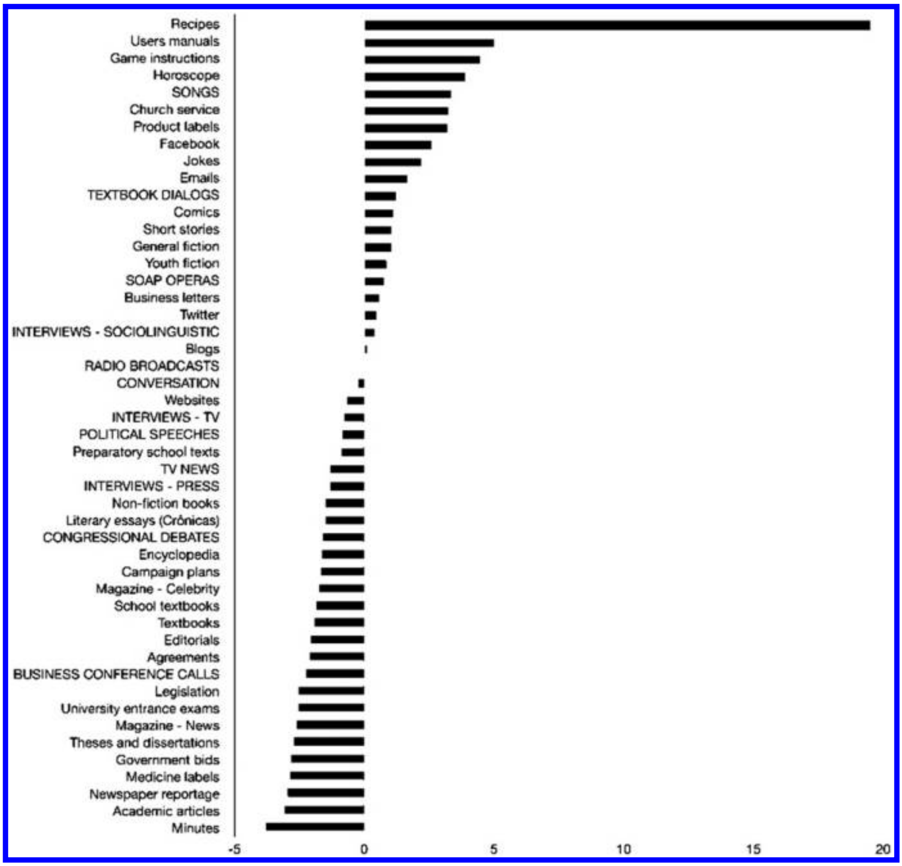

Figure 5: Dimension 4: Directive Discourse

Modo de Preparo: *Lave* o *limão*, *corte* em quatro *pedaços*. *Coloque* em um *copo* baixo e largo, *junte açúcar* e, com um *socador*, *amasse* até liberar todo o *suco*. *Acrescente gelo* triturado, *pinga*, *misture bem* e *sirva*.

[*Capirinha*
1 lime
1 tablespoon sugar
2 crushed ice cubes
4 tablespoons *cachaça*

Preparation: Wash the lime, cut it in four wedges. Put in a low wide glass, add the sugar, and with a crusher, mash until the juice from the lime has been squeezed out. Add the crushed ice, *cachaça*, mix well and serve.]

The fact that medicine labels (*bulas*) are not marked for this dimension runs counter to expectations, as experience with such texts would suggest that they are directive. However, both the present subjunctive and the imperative – the main features marking directiveness on Dimension 4 – have low means in this register, (4.1 and 1.2 per thousand words, respectively), compared to the most marked registers (namely, 39.3 and 35.3 for recipes, and 21.1 and 14.3 for owner's manuals; the corpus means are respectively 2.05 and 3.5, see Table 2). A close examination of individual texts shows that directiveness in medicine labels in Brazilian Portuguese is typically expressed by features such as the infinitive form (*aplicar*, etc.), the modal *dever* (usually in the negative), the adverb *somente* (only), and directly by numbers indicating the recommended dosage levels. The sample in Example 6 illustrates the use of these features in a *bula*.

(*6*) Dipirona *não deve* ser administrada em altas doses ou por períodos prolongados, sem controle médico. POSOLOGIA E ADMINISTRAÇÃO. Criança de 5,5 a 7,5 Kg: *0,1 à 0,2* ml - *somente* intramuscular. Criança de 8 a 10 Kg: *0,1 à 0,3* ml - *somente* intramuscular. (. . .) Adultos e adolescentes acima de 15 anos: *2 à 5 ml* - IM ou IV. (. . .) Doses maiores, *somente* à [sic] critério médico. *Aplicar* a injeção endovenosa lentamente, 1 ml/minuto. Não *misturar* medicamentos na mesma seringa. PRECAUÇÕES. O uso de Dipirona *deve* ser evitado nos três primeiros meses e nas últimas 6 semanas da gestação e, mesmo fora destes períodos, *somente administrar* em gestantes em casos de extrema necessidade.

[Dipyrone should not be administered in large doses or for prolonged periods without medical supervision. DOSAGE AND ADMINISTRATION. Children 5.5 to 7.5 kg: 0.1 to 0.2 ml—intramuscularly only. Children 8 to 10 kg: 0.1 to 0.3 ml—intramuscularly only. (. . .) Adults and adolescents over 15 years: 2 to 5 ml—IM or IV. (. . .) Larger doses, only by prescription. Apply intravenous injection slowly, 1 ml/minute. Do not mix medications in the same syringe. PRECAUTIONS. The use of dipyrone should be avoided in the first three months and the last 6 weeks of pregnancy, and even outside of these periods, only use in pregnant women in cases of extreme necessity.]

Factor 5 has fifteen features (Table 7), with the main ones being the future subjunctive and the present future indicative – both of which mark a time orientation toward the future. *Ou* (or) co-ordination can link both clauses and phrases, creating sequences of independent clauses or phrasal elements. The modal *dever* can be used as an obligation or a prediction modal, and *poder* can act as a prediction, ability or permission modal. Subordinating conjunctions create dependent clauses, which in turn are attached to the main clause in different ways, such as through condition, cause or comparison relationships. Probability adverbs mark the likelihood of an action or state.

Feature [label]	Loading
Verbs: Future subjunctive [vbsubfut]	0.616
Conjunctions: Co-ordinating (*ou*) [cjou]	0.611
Verbs: Future present tense [vbfutpres]	0.513
Modals: *Dever* [mddever]	0.474
Modals: *Poder* [mdpoder]	0.426
Subordinating (conditional) clause [cjcond]	0.390
Adverbs: Likelihood [advlikl]	0.389
Conjunctions: Co-ordinating (phrasal) [cjcoorphr]	0.322
(Adjectives: Relational [adjrela]	0.303)
Nouns: Place [nplac]	−0.301
Verbs: Past subjunctive [vbsubpast]	−0.308
(Articles: Indefinite [artindef]	−0.320)
Adjectives: Affiliative [adjaffi]	−0.355
Verbs: Imperfect [vbimpf]	−0.375
Verbs: Past indicative tense [vbpast]	−0.554

Table 7: Inter-factor correlations

Phrasal co-ordination enables the linking of nouns and adjectives in phrases. Relational adjectives are classifiers that 'have little descriptive content' (Biber *et al*., 1999: 508). In contrast, the main features on the negative pole of the factor mark a time orientation toward the past: the past indicative and the past imperfect tenses as well as the past subjunctive. In addition to these, affiliative adjectives express local and national designations, indefinite articles narrow down 'the reference to a single member of a class' (Biber *et al*., 1999: 70), and place nouns are used to refer to particular places in the discourse. These two poles basically mark a distinction between future and past; therefore, the interpretive label 'Future *versus* past time orientation' is used for Dimension 5.

The distribution of registers on Dimension 5 is shown in Figure 6. At the top of the chart are the registers using primarily the future tense in conjunction with *dever* and *poder* modals to express conditions, abilities and obligations. Example 7 illustrates the use of such linguistic characteristics, including *poder* (*poderão*), future subjunctive (*se a empresa … se fizer*), future present (*reger-se-á*, etc.), *ou* conjunctions (*licitar ou contratar*) and phrasal co-ordination (*normas e procedimentos*, etc.), in a government bid solicitation.

(7) A proponente que chegar atrasada *será* desclassificada. A presente licitação *reger-se-á* pelas normas *e* procedimentos do Regulamento Interno de Licitações e Contratos, *e* pelo presente instrumento convocatório. Participação: *Poderão* participar desta licitação empresas que comprovem estar devidamente regularizadas perante os órgãos públicos competentes (…). Não *poderão* participar da

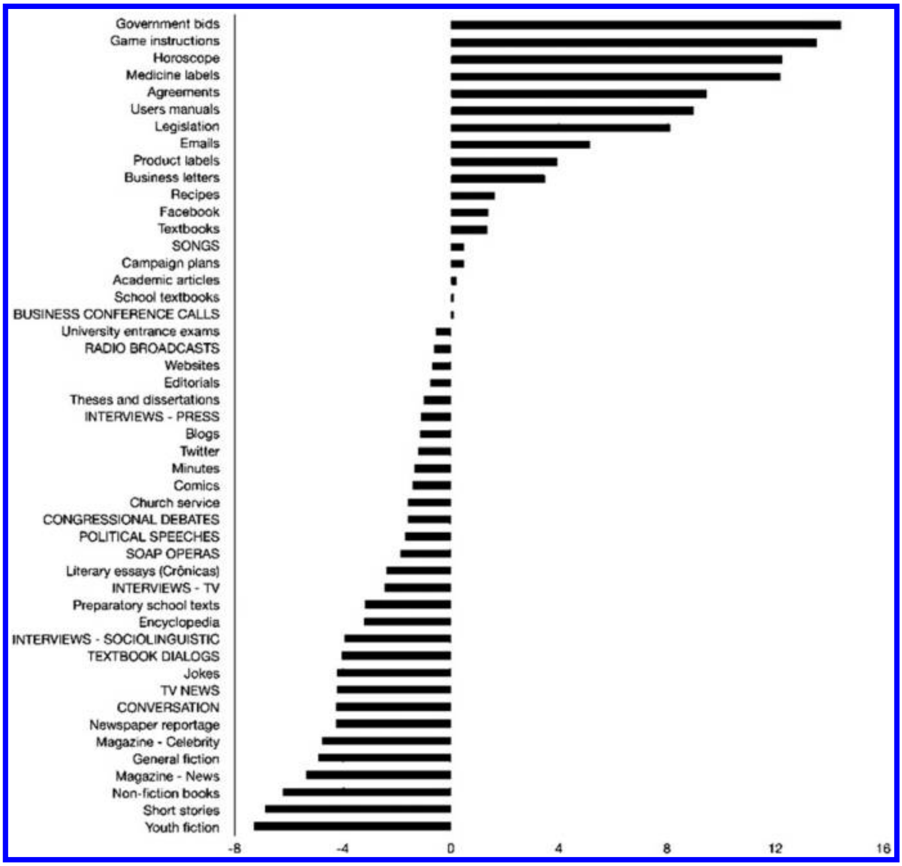

Figure 6: Dimension 5: Future *versus* Past Time Orientation

> presente licitação: Empresas suspensas de licitar *ou* contratar com o [Government office] (. . .). *Se* a empresa licitante *se fizer* representar pela sua filial, para atender o objeto do presente edital, deverá também apresentar todos os documentos elencados (. . .)
>
> [All late bidders will be disqualified. The present solicitation shall be governed by the provisions of the Internal Bids and Contracts Code and by the present solicitation instrument. Participation: Participation shall be granted to all businesses that demonstrate to be in good standing with the relevant public offices (. . .). The following shall not be able to participate in the current bid: All businesses that have been suspended from bidding or contracting with [Government office] (. . .). Should the bidding company be represented by its subsidiary, so as to meet the object of this notice, it must present all the required documents (. . .)]

At the other extreme are registers that rely on past-tense forms to mark narrative concerns – primarily fiction registers, but also including other forms

Feature [label]	Loading
Pronouns: Rare in object position [objprnrare]	0.628
Verbs: Second person [vb2]	0.466
Pronouns: Possessive [prnposs]	0.424
Subordinating (final) clause [cjfinal]	0.413
(*Que* clause controlled by preposition [clqueeprp]	0.380)
Pronouns: Third person, object position [prn3obl]	0.371
(Pronouns: Relative *que* [prnque]	0.340)
Verbs: Public [vbpubl]	0.327
Modals: *Haver que / haver de* (have to, ought to) [mdhaver]	0.311
(Adjectives: Evaluative [adjeval]	−0.318)
(Yes or no question [qsyn]	−0.330)
(Adverbs: Amplifier [advampl]	−0.340)
(Adverbs: Intensity [advints]	−0.341)

Table 8: Loadings on Dimension 6: Reported Discourse. (*Note*: Features with larger weights on a different factor are enclosed in parentheses)

of storytelling, such as magazine news, newspaper reportage, conversation and jokes. Example 8 illustrates the use of past tense forms (*vivia*, etc.) in an excerpt from a youth fiction story.

(*8*) A Marcella *vivia* me protegendo, desde pequeno. Talvez porque, quando eu *era* bem criança, *tive* bronquite alérgica. Nem lembro bem como *era*, mas dizem que eu *tossia* tanto que até *tinham* medo de que eu *botasse* o pulmão pra fora. Desde então, ela *cuidava* de mim. Sempre me *ajudava* nos trabalhos da escola.

[Marcella would always protect me, ever since I was little. Maybe because when I was a small kid, I had allergic bronchitis. I don't remember what it was like, but they say I would cough so much that they were afraid I would cough my lungs out. Since then, she has always taken care of me. Always helped me with my homework.]

Dimension 6 (see Table 8) has in fact only one pole (the positive one) because all features loading on the negative pole have higher loadings on other factors. Rare personal pronouns in object position include the formal and archaic (in Brazilian Portuguese) *vos* (second-person plural) and *los* (third-person plural) as well as contractions between some of these pronouns and preposition, such as *convosco*, formed by joining the preposition *com* ('with') with the pronoun *vos*. Possessive pronouns refer to the participants in the text, and second-person verb forms are used to

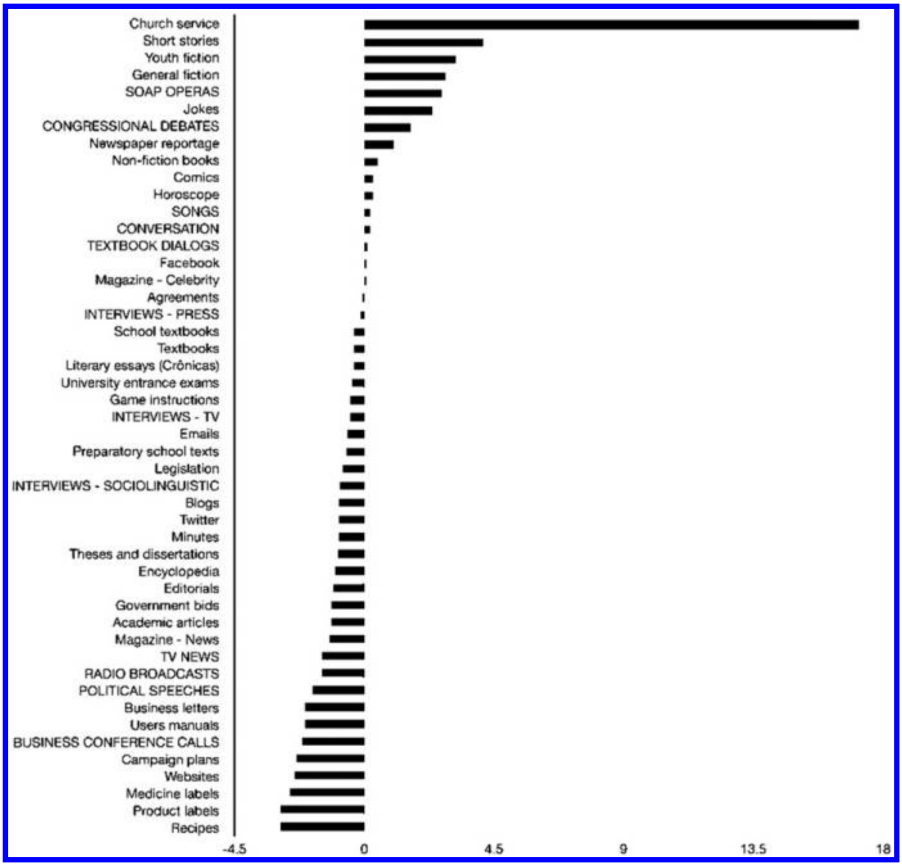

Figure 7: Dimension 6: Reported Discourse

refer to an addressee; therefore, both are related to interactivity (Biber, 1988: 105). Final subordinating conjunctions (e.g., *para que*, 'so that') are complex subordinators (Biber *et al*., 1999: 282) that mark purpose in dependent clauses (Quirk *et al*., 1985: 1070). *Que* clauses controlled by a preposition perform a number of different functions, depending on the preposition attached to them; in general, these are considered markers of formal discourse. Third-person pronouns in the object position refer to human or non-human referents in the discourse. *Que* relative clauses, as previously mentioned, provide ways in which to make elaborated reference to discourse participants. Public verbs can 'function as markers of indirect, reported speech' (Biber, 1988: 109). Finally, the modal *haver* can be used as a formal marker of obligation. Based on the functions shared by these co-occurring features, the interpretive label for this dimension is 'Reported Discourse'.

The distribution of registers on Dimension 6 is plotted in Figure 7, which shows that the most marked register is church services, followed by fiction and different story-telling registers in a distant second place. All of these registers rely on public verbs to report speech, but Order of Mass texts also make dense use of rare oblique pronouns and second-person verb forms (which are very infrequent in the other registers), thereby boosting their dimension scores.

Example 9 is an excerpt of the Eucharistic prayer from an Order of Mass leaflet that illustrates the frequent use of second-person verb forms (*tomai*, etc.), normally in the imperative, in addition to public verbs (*dizendo*, *etc.*), possessive pronouns[6] (*seus*, etc.) and the archaic *vossa* object pronoun.

(*9*) Ele tomou o cálice em *suas* mãos, deu graças novamente, e o deu a *seus* discípulos, *dizendo*: *Tomai*, todos, e *bebei*: este é o cálice do *meu* sangue, o sangue da nova e eterna aliança, que será derramado por *vós* e por todos, para remissão dos pecados. *Fazei* isto em memória de mim. Eis o mistério da fé! Todas as vezes que comemos deste pão e bebemos deste cálice, *anunciamos*, Senhor, a *vossa* morte, enquanto esperamos a *vossa* vinda!

[Jesus took the chalice, and gave it to the disciples, and said, ‘Take this, all of you, and drink from it; for this is the chalice of my blood, the blood of the new and everlasting covenant, which will be poured out for you and for many for the forgiveness of sins. Do this in memory of me. The mystery of faith! When we eat this Bread and drink this Cup, we proclaim your Death, O Lord, until you come again.]

To summarise, the following dimensions were identified for Brazilian Portuguese:

(*1*) Oral *versus* Literate Discourse
(*2*) Argumentation
(*3*) Involved *versus* Informational Production
(*4*) Directive discourse
(*5*) Future *versus* Past Orientation
(*6*) Reported Discourse

The F-test (ANOVA) performed on the mean dimension scores on each dimension indicates whether significant differences exist among the registers with respect to their mean dimension scores. In addition, the R^2 statistic ‘measures the percentage of the variance among dimension scores that can be predicted by knowing the register categories’ (Biber, 1995: 119). Both statistics are shown in Table 9. The ANOVA results indicate that the differences among the registers are significant on all dimensions; the R^2

[6] These are referred to as possessive pronouns and not possessive determiners in Brazilian Portuguese grammars (e.g., Bechara, 1999) and in Biber (1988: 214).

Dimension	F	p	df	R^2 (percent)
1	97.037	.000	46	83.3
2	19.926	.000	46	48.1
3	101.521	.000	46	83.7
4	57.961	.000	46	73.5
5	4.561	.000	46	67.6
6	43.525	.000	46	69.2

Table 9: Inter-factor correlations

values suggest that five dimensions (all except Dimension 2) can be regarded as strong predictors of register differences in Brazilian Portuguese. These values are comparable to previous studies, such as those for English (Biber, 1988: 182), in which R^2 values ranged from 16.9 to 84.3; Tuvaluan, with values from 30 to 70; Korean, 5.5 to 61.5; and Somali, 19.4 to 90.9 (all from Biber, 1995).

4. Discussion and conclusion

Previous language-wide MD studies have all identified a dimension that reflects the distinction between oral and literate discourse, generally as the first and, thus, main factor. In English, it was initially called 'Interactive versus Edited Text' (Biber, 1986) but was later renamed as 'Involved versus Informational Production' (Biber, 1988). In Nukulaelae Tuvaluan, it is 'Interpersonal versus Information Reference' Dimension 2 (Biber, 1995), in Korean, 'On-line Interaction versus Planned Exposition' Dimension 1 (Biber, 1995), in Somali, 'Structural Elaboration: Involvement versus Exposition' Dimension 1 (Biber, 1995), and in Spanish,'Oral versus Literate Discourse' (Biber *et al.*, 2006; and Biber and Tracy-Ventura, 2007). Both of our Dimensions 1 and 3 reflect this potentially universal dimension. Our Dimension 2 (Argumentation) occurs in English (Dimension 4) and in Somali in Dimensions 3 and 6 (Biber, 1995). Our Dimension 4 (Directive Discourse) appears in Biber's (2006) study of university language and in Friginal's (2009) research on call centres (in both as Dimension 2). It so happens that our Dimension 5 (Future *versus* Past Orientation) has no direct parallel, as it marks a distinction between future and past, not present and past, which would be more natural for narrativity. Finally, a 'reporting dimension' such as our Dimension 6 is found in other MD studies, including the early MD analysis of English (Dimension 3: 'Reported versus Immediate Style'; Biber, 1986), Korean (Dimension 5) and Somali (Dimension 3).

At the same time, two particular characteristics distinguish our dimensions from previous MD research. The first is that our dimensions include both oral *versus* literature (Dimension 1), and involved *versus* informational (Dimension 3), which in other studies were regarded as just two different labels for the same parameter. The inter-factor correlation between these two dimensions is the highest (0.41) in this study, thereby reflecting their mutual relationship. The second distinguishing characteristic is that it did not reveal a clear-cut narrative dimension. This lack of a standard narrative dimension might have been the result of our corpus design, in which the representation of literature (youth fiction, general fiction and short stories) is lower than in other MD studies. The share of fiction in the written component of the corpora in previous MD research ranges from 7.5 percent in Spanish (Biber *et al*., 2006: 7) to 12.5 percent in Korean (Biber, 1995: 91), 21.7 percent in English (Biber, 1988: 67), and 6.3 percent in ours. In addition, the uniqueness of our Dimension 5 is derived from the unique set of registers present in the corpus – many of which have marked frequencies for future tenses, such as government bids, game instructions, horoscopes and drug labels (*bulas*). As these registers are not usually found in other MD studies, a similar dimension has not been replicated. This is a reminder that MD studies differ 'with respect to the set of linguistic features included in the analysis, and the set of registers represented in the corpus for analysis'; therefore, 'it would be reasonable to expect that the parameters of variation that emerge from each analysis would be fundamentally different' (Biber, 2013b: 145). In short, 'strictly speaking, dimensions are valid only for the corpus that they are derived from' (Biber, 2013b: 148).

As mentioned above, the only other MD study on Brazilian Portuguese to date is Kauffmann (2005), whose dimensions generally match those found in this study. The narrative end of his first dimension is similar to our Dimension 5, as both mark past processes through the common use of past tense and imperfect preterite verbs. Newspaper registers marked for narrativity in Kauffmann, like newspaper reportage, press interviews and essays (*crônicas*), are all included on the past orientation end of our Dimension 5 as well. In this way, our study corroborates the notion that a time orientation toward the past is central to the characterisation of particular newspaper registers. Kauffmann's second dimension, marking argumentation, is similar to our Dimension 2 with respect to the register distinctions: in both studies, press interviews, essays (*crônicas*) and editorials are on the argumentative pole of the dimension. However, the two dimensions share little in common with respect to the features used to mark argumentation. In Kauffmann's study, the present indicative, demonstrative pronouns and word count were the features that loaded on the argumentative pole of the dimension. Of these, only demonstrative pronouns are present in our Dimension 2; the bulk of our Dimension 2 consists of *que*- and infinitive clauses, which were not included in Kauffmann's analysis. Unlike Kauffmann (2005), which also employed the PALAVRAS parser, the current study made use of a post-processor program that combined features tagged by PALAVRAS to derive additional characteristics, such as the different complement clauses in Brazilian Portuguese. However, interestingly, both studies arrived at similar register characterisations of the same language based on largely different linguistic feature pools. One reason for this is that the interpretive labels given to the dimensions are motivated by the consideration of the communicative functions at play in whole texts, and not simply by what could be generalised from the functions enacted by the linguistic features loaded on the factor. Thus, different analysts can legitimately arrive at similar dimensions from different feature sets. This highlights the major influence of text interpretation in uncovering the underlying communicative processes signalled by the dimensions – a qualitative aspect of the MD toolbox that is not immediately apparent to non-analysts.

Portuguese and Spanish are two closely related languages, and a comparison of our findings with those of previous MD analyses of Spanish (Biber *et al.*, 2006; and Biber and Tracy-Ventura, 2007) shows that, despite both analyses having identified the same number of dimensions (six) for each language, only two common dimensions surfaced in both languages: oral/literate discourse (Dimension 1 in both studies) and narrative discourse (Dimension 3 in Spanish and Dimension 5 here, with the caveat previously discussed). This suggests that, although the two languages have many linguistic features in common (notably, the vocabulary), they use linguistic resources differently for communicative purposes. Again, part of the mismatch might be attributable to differences in corpus design, as discussed here. However, as Biber *et al.* (2006: 30) argued, we should not overstate this influence if the corpora 'cover roughly the same range of registers, differing primarily in the relative weightings given to particular registers'. More research is needed to assess the influence of different corpora on MD analyses to resolve this issue.

The findings reported here could have implications for the teaching of Brazilian Portuguese in both L1 and foreign language contexts. For instance, the features marking each dimension are routinely included in most curricula, but are normally taught from a purely structural point-of-view. However, the dimensions show that they perform both individual and shared functions and, as such, could be taught with a focus on discourse and communication, with text samples taken from those registers where they are most active.

Although other MD studies, such as for English and Spanish, opted to generalise the findings to different national varieties of those languages, consideration of the marked syntactic and lexical differences among Brazilian, European, and African varieties of Portuguese (Berber Sardinha and São Bento Ferreira, 2014; and Castilho, 2009) suggested that it would be prudent to focus on the Brazilian variety only. Future MD research should consider exploring other varieties of Portuguese to determine if there are indeed significant differences in the multidimensional space of register variation in the major dialects of Portuguese around the world.

Acknowledgments

The authors wish to thank both Fapesp (the São Paulo State Research Council) and CNPq (Brasília, DF) for the grants that supported the Dimensions of Register Variation in Brazilian Portuguese project. We are also grateful to the two anonymous reviewers for their careful reading and insightful comments.

References

Atkinson, D. 2001. 'Scientific discourse across history: a combined multidimensional/rhetorical analysis of the Philosophical Transactions of the Royal Society of London' in S. Conrad and D. Biber (eds) Variation in English: Multi-Dimensional Studies, pp. 45–65. Harlow: Longman.

Azevedo, M.M. 2005. Portuguese: A Linguistic Introduction. Cambridge: Cambridge University Press.

Bacelar do Nascimento, M.F., A. Mendes, S. Antunes and L. Pereira. 2014. 'The Reference Corpus of Contemporary Portuguese and related resources' in T. Berber Sardinha and T. São Bento Ferreira (eds) Working with Portuguese Corpora, pp. 237–56. London and New York: Bloomsbury/Continuum.

Bechara, E. 1999. Moderna Gramática Portuguesa [Modern Portuguese grammar]. (Thirty-seventh edition.) Rio de Janeiro: Lucerna.

Berber Sardinha, T. (ed.). 2005. A língua Portuguesa no Computador [The Portuguese Language on Computer]. Campinas, São Paulo: Mercado de Letras/FAPESP.

Berber Sardinha, T. 2014. '25 years later: comparing Internet and pre-Internet registers' in T. Berber Sardinha and M. Veirano Pinto (eds) Multi-Dimensional Analysis 25 Years On: A Tribute to Douglas Biber, pp. 35–80. Amsterdam and Philadelphia: Johns Benjamins.

Berber Sardinha, T. and T. São Bento Ferreira. 2014. Working with Portuguese Corpora. London and New York: Bloomsbury / Continuum.

Berber Sardinha, T., T. São Bento Ferreira and R. d. B. S. Teixeira. 2014. 'Lexical bundles in Brazilian Portuguese' in T. Berber Sardinha and T. São Bento Ferreira (eds) Working with Portuguese Corpora, pp. 33–68. London and New York: Bloomsbury/Continuum.

Bértoli Dutra, P. 2014. 'Multi-Dimensional Analysis of pop songs' in T. Berber Sardinha and M. Veirano Pinto (eds) Multi-Dimensional Analysis 25 Years On: A Tribute to Douglas Biber, pp. 151–81. Amsterdam and Philadelphia: Johns Benjamins.

Besnier, N. 1988. 'The linguistic relationships of spoken and written Nukulaelae registers', Language 64 (4), pp. 707–36.

Biber, D. 1985. 'Investigating macroscopic textual variation through multifeature/multidimensional analyses', Linguistics 23 (2), pp. 337–60.

Biber, D. 1986. 'Spoken and written textual dimension in English: resolving the contradictory findings', Language 62 (2), pp. 384–414.

Biber, D. 1988. Variation across Speech and Writing. Cambridge: Cambridge University Press.

Biber, D. 1995. Dimensions of Register Variation: A Cross-linguistic Comparison. Cambridge: Cambridge University Press.

Biber, D. 2000. 'Investigating language use through corpus-based analyses of association patterns' in M. Barlow and S. Kemmer (eds) Usage-based Models of Language, pp. 287–313. Stanford: Center for the Study of Language and Information.

Biber, D. 2006. University Language: A Corpus-based Study of Spoken and Written Registers. Amsterdam and Philadelphia: John Benjamins.

Biber, D. 2012. 'Register as a predictor of linguistic variation', Corpus Linguistics and Linguistic Theory 8 (1), pp. 9–37.

Biber, D. 2013a. 'Interview with Douglas Biber [conducted by Bethany Gray]', Journal of English Linguistics 4 (4), pp. 359–79.

Biber, D. 2013b. 'Twenty-five years of Biber's Multi-Dimensional Analysis: introduction to the special issue and an interview with Douglas Biber [conducted by Eric Friginal]', Corpora 8 (2), pp. 137–52.

Biber, D. 2014. 'Multi-Dimensional Analysis: a personal history' in T. Berber Sardinha and M. Veirano Pinto (eds) Multi-Dimensional Analysis 25 Years On: A Tribute to Douglas Biber, pp. xxix–xxxviii. Amsterdam and Philadelphia: Johns Benjamins.

Biber, D. and S. Conrad. 2009. Register, Genre and Style. Cambridge: Cambridge University Press.

Biber, D., M. Davies, J.K. Jones and N. Tracy-Ventura. 2006. 'Spoken and written register variation in Spanish: a multi-dimensional analysis', Corpora 1 (1), pp. 1–37.

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., S. Johansson, G. Leech, S. Conrad and E. Finegan. 1999. The Longman Grammar of Spoken and Written English. London: Longman.

Biber, D. and J. Kurjian. 2007. 'Towards a taxonomy of web registers and text types: a multi-dimensional analysis' in M. Hundt, N. Nesselhauf and C.

Biewer (eds) Corpus Linguistics and the Web, pp. 109–32. Amsterdam and New York: Rodopi.

Biber, D. and N. Tracy-Ventura. 2007. 'Dimensions of register variation in Spanish' in G. Parodi (ed.) Working with Spanish Corpora, pp. 54–89. London: Continuum.

Bick, E. 2014. 'PALAVRAS, a constraint grammar-based parsing system for Portuguese' in T. Berber Sardinha and T. São Bento Ferreira (eds) Working with Portuguese Corpora, pp. 279–302. London and New York: Bloomsbury/Continuum.

Carroll, J.B. 1960. 'Vectors of prose style' in T.A. Sebeok (ed.) Style in Language, pp. 283–92. Cambridge, Massachusetts: Technology Press of Massachusetts Institute of Technology.

Castilho, A.T. d. 2009. 'Portuguese' in K. Brown and S. Ogilvie (eds) Concise Encyclopedia of Languages of the World, pp. 883–5. Oxford: Elsevier.

Castilho, A.T. d. (ed.). 1989. Português Culto Falado no Brasil [Standard Spoken Portuguese in Brazil]. Campinas, São Paulo: Editora da Unicamp.

Conde, H.M. d. A. 2002. Escolhas Léxico-Gramaticais em Composições de Alunos Avançados de Inglês Originários de Instituições de Ensino Bilíngues e Monolíngües – Um Estudo Multidimensional Baseado em Corpus [Lexico-Grammatical Choices in Advanced Student Writing From Bilingual and Monolingual Schools – A Multidimensional Corpus-based Study]. Unpublished MA thesis. Brazil: São Paulo Catholic University, São Paulo.

Condi de Souza, R. 2014. 'Dimensions of variation in TIME magazine' in T. Berber Sardinha and M. Veirano Pinto (eds) Multi-Dimensional Analysis 25 Years On: A Tribute to Douglas Biber, pp. 179–98. Amsterdam and Philadelphia: Johns Benjamins.

Conrad, S. 2014. 'Expanding Multi-Dimensional Analysis with qualitative research techniques' in T. Berber Sardinha and M. Veirano Pinto (eds) Multi-Dimensional Analysis 25 Years On: A Tribute to Douglas Biber, pp. 275–98. Amsterdam and Philadelphia: Johns Benjamins.

Crossley, S. and M.M. Louwerse. 2007. 'Multi-dimensional register classification using bi-grams', International Journal of Corpus Linguistics 12 (4), pp. 453–78.

Crossley, S., L.K. Varner and D. McNamara. 2014. 'A Multi-Dimensional Analysis of essay writing: what linguistic features tell us about situational parameters and the effects of language functions on judgments of quality' in T. Berber Sardinha and M. Veirano Pinto (eds) Multi-Dimensional Analysis 25 Years On: A Tribute to Douglas Biber, pp. 199–240. Amsterdam and Philadelphia: Johns Benjamins.

Cunha, C. 2001. Nova Gramática do Português Contemporâneo [New Grammar of Contemporary Portuguese]. Rio de Janeiro, Brazil: Nova Fronteira.

Delegá-Lúcio, D. 2013. A Variação Entre Textos Argumentativos e o Material Didático de Inglês: Aplicações sa Anaçlise Multidimensional e do Corpus Internacional de Aprendizes de Inglês (ICLE) [Variation Across Argumentative Texts in The Context of Designing English Teaching Materials: A Multi-Dimensional Analysis of The International Corpus of Learner English (ICLE)]. Unpublished MA thesis. Brazil: São Paulo Catholic University, São Paulo.

Friginal, E. 2009. The Language of Outsourced Call Centers: A Corpus-based Study of Cross-cultural Interaction. Amsterdam and Philadelphia: John Benjamins.

Grieve, J., D. Biber, E. Friginal and T. Nekrasova. 2010. 'Variation among blogs: a multi-dimensional analysis' in A. Mehler, S. Sharoff and M. Santini (eds) Genres on the Web: Computational Models and Empirical Studies, pp. 303–22. Dordrecht and New York: Springer.

Ilari, R. 1991. Gramática do Português Falado: Níveis de Análise Linguística [Grammar of Spoken Portuguese: Levels of linguistic Analysis]. Campinas, São Paulo: Editora da Unicamp.

Kauffmann, C. 2005. O Corpus do Jornal: Variação Linguística, Gêneros e Dimensões da Imprensa Diária Escrita [The Newspaper Corpus: Linguistic Variation, Genres, and Dimensions in The Daily Press]. Unpublished MA thesis. Brazil: São Paulo Catholic University, São Paulo.

Kilgariff, A., M. Jakubíček, J. Pomikalek, T. Berber Sardinha and P. Whitelock. 2014. 'PtTenTen: a corpus for Portuguese lexicography' in T. Berber Sardinha and T. São Bento Ferreira (eds) Working with Portuguese Corpora, pp. 111–30. London and New York: Bloomsbury/Continuum.

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.

Lamb, W. 2008. Scottish Gaelic Speech and Writing: Register Variation in an Endangered Language. Belfast: Cló Ollscoil na Banríona.

Lee, D.Y.W. 1999. Modelling Variation in Spoken and Written Language: The Multi-Dimensional Approach Revisited. Unpublished PhD thesis. Lancaster University, Lancaster, UK.

de Mönnink, I.M., N. Brom and N.H.J. Oostdijk. 2003. 'Using the MF/MD method for automatic text classification' in S. Granger and S. Petch Tyson (eds) Extending the Scope of Corpus-based Research: New Applications, New Challenges, pp. 15–25. Amsterdam: Rodopi.

Moura Neves, M.H. de. 2000. Gramática de Usos do Português [Portuguese Usage Grammar]. São Paulo: Editora Unesp.

Parodi, G. 2007. 'Variation across registers in Spanish: exploring the El-Grial PUCV Corpus' in G. Parodi (ed.) Working with Spanish Corpora, pp. 11–53. London: Continuum.

Preti, D. (ed.). 2005. O Discurso Oral Culto [Standard Oral Discourse]. São Paulo: Humanitas.

Quaglio, P. 2009. Television Dialogue: The Sitcom Friends vs. Natural Conversation. Amsterdam and Philadelphia: John Benjamins.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik. 1985. A Comprehensive Grammar of the English Language. London: Longman.

Reppen, R. 2001. 'Register variation in student and adult speech and writing' in S. Conrad and D. Biber (eds) Variation in English: Multi-Dimensional Studies, pp. 187–99. Harlow: Longman.

Shergue, O. 2003. Dimensão de Variação no Discurso Médico-Acadêmico: O Artigo de Pesquisa e a Apresentação de Trabalhos Científicos em Congressos [Dimensions of Variation In Medical Discourse in Academia: Research Articles and Conference Paper Presentations]. Unpublished MA thesis. Brazil: São Paulo Catholic University, São Paulo.

Souza e Silva, M.C.P. de and B. Brait (eds). 2013. Texto ou Discurso? [Text or Discourse?]. São Paulo: Contexto.

Titak, A. and A. Roberson. 2013. 'Dimensions of web registers: an exploratory multi-dimensional comparison', Corpora 8 (2), pp. 235–60.

Thomas, E.W. 1969. The Syntax of Spoken Brazilian Portuguese. Nashville, Tennessee: Vanderbilt University Press.

Veirano Pinto, M. 2014. 'Dimensions of variation in North American movies' in T. Berber Sardinha and M. Veirano Pinto (eds) Multi-Dimensional Analysis 25 Years on: A Tribute to Douglas Biber, pp. 109–50. Amsterdam and Philadelphia: Johns Benjamins.

Whitlam, J. 2010. Modern Brazilian Portuguese Grammar. New York: Routledge.

Zuppardo, M.C. 2014. Dimensões de Variação em Manuais Aeronáuticos: Um Estudo Baseado na Análise Multi-Dimensional [Dimensions of Variation in Aviation Manuals: A Mutli-Dimensional Study]. Unpublished MA thesis. Brazil: São Paulo Catholic University, São Paulo.

Appendix A: Composition of the Brazilian Register Variation Corpus (CBVR)

	Register*	Words	Percent of total no. of words
1	Academic articles	92,148	1.6
2	Agreements (*a*)	44,562	0.8
3	Blogs	31,486	0.6
4	BUSINESS CONFERENCE CALLS	106,076	1.9
5	Business letters	12,720	0.2
6	Campaign plans	29,724	0.5
7	Church service	66,995	1.2
8	Comics (*b*)	25,937	0.5
9	CONGRESSIONAL DEBATES (*c*)	641,080	11.4
10	CONVERSATION	93,470	1.7
11	Editorials	11,233	0.2
12	Emails – Personal	11,223	0.2
13	Encyclopedia entries	13,690	0.2
14	Essays (*d*)	21,403	0.4
15	Facebook	11,022	0.2
16	Game instructions (*e*)	16,260	0.3
17	General fiction	403,796	7.2
18	Government bids (*f*)	127,239	2.3
19	Horoscope	12,637	0.2
20	INTERVIEWS – SOCIOLINGUISTIC	152,788	2.7
21	INTERVIEWS – PRESS	28,175	0.5
22	INTERVIEWS TV	263,821	4.7
23	Jokes (*g*)	9,310	0.2
24	Legislation	125,531	2.2
25	Magazine celebrity	25,738	0.5
26	Magazine news	19,850	0.4
27	Medicine / drug labels	16,061	0.3
28	Minutes (*h*)	25,929	0.5
29	Newspaper reportage	11,467	0.2
30	Non-fiction books	55,028	1.0
31	POLITICAL SPEECHES	44,591	0.8
32	Prep. school texts (*i*)	15,411	0.3
33	Product labels	9,183	0.2

Appendix A (*Continued*): Composition of the Brazilian Register Variation Corpus (CBVR)

	Register*	Words	Percent of total no. of words
34	RADIO BROADCASTS	91,335	1.6
35	Recipes	9,591	0.2
36	Short stories	57,362	1.0
37	SOAP OPERAS	93,627	1.7
38	SONGS	11,990	0.2
39	TEXTBOOK DIALOGUES (*j*)	9,447	0.2
40	Textbook texts (*k*)	12,732	0.2
41	Textbooks (*l*)	1,234,790	21.9
42	Theses	617,943	10.9
43	TV NEWS	11,453	0.2
44	Twitter	11,027	0.2
45	User's / owner's manuals	301,650	5.3
46	Websites	28,338	0.5
47	Written exams	33,937	0.6
48	Youth fiction	543,200	9.6
	Total	5,644,006	100.0

Key

*	Registers in UPPER CASE are spoken.
(*a*)	Legal documents dealing with services, sales, rentals, *etc*.
(*b*)	Comic books for children and adolescents
(*c*)	Verbatim records of debates and discussions in the Brazilian Congress
(*d*)	Short fictional/narrative pieces, generally printed in newspapers (*crônicas*)
(*e*)	Instruction leaflets that accompany board games
(*f*)	Government contract announcements
(*g*)	Funny stories, with a punch line, from specialised websites
(*h*)	Records of meetings of different kinds (apartment building tenants, NGOs, accident prevention committees, *etc*.)
(*i*)	Texts from university entrance exam practice manuals (*apostilas de cursinho*)
(*j*)(*k*)	Texts from a Portuguese as a foreign language course book
(*l*)	University-level books, each on a different subject (business administration, philosophy, psychology, *etc*.)