Dimensions of variation in spoken Brazilian Portuguese

Tony BERBER SARDINHA (PUC-SP)
Carlos KAUFFMAN (Folha de S. Paulo News Organization)
Cristina MAYER ACUNZO (PUC-SP)

Multi-Dimensional Analysis is a methodology introduced by Biber (1988, inter alia) that allows for the identification of underlying parameters of variation in corpus data, typically across different registers. Dimensions of variation, in turn, are patterns of cooccurrence of linguistic features underlying the registers of a language. An example of dimension of variation (for English) is 'Interaction versus Information' (Biber, 1988), which maps, along a scale, how different registers are more or less interactive or more or less informational. As the name implies, Multi-Dimensional Analysis typically reveals a multitude of dimensions, each representing a scale of variation. Multi-Dimensional Analysis makes extensive use of statistical techniques, notably Factor Analysis, for the extraction of factors that are then interpreted both linguistically and situationally to indicate dimensions of variation. Previous research includes analysis of both whole languages and individual registers. Examples of the former are the descriptions of English (Biber, 1988; Crossley & Louwerse, 2007; de Mönnink, et al., 2003; Lee, 1999), Korean (Kim & Biber, 1994), Somali (Biber & Hared, 1994), Nukulaelae (Besnier, 1988), Gaelic (Lamb, 2008) and Spanish (Biber, et al., 2006; Parodi, 2007); examples of the latter are analyses of conversation (Biber, 2004), sitcoms (Quaglio, 2009) and research articles (Biber, et al., 1994). In this paper, we present a synchronic study of register variation in Brazilian Portuguese. Portuguese is an important European language, the second largest Romance language, and the Brazilian variety accounts for 90% of its native speakers. To date, no Multi-Dimensional Analysis has been carried out on Portuguese. In this paper, we present a Multi-Dimensional Analysis of different spoken registers (ranging from face-to-face conversation to radio, TV shows, and telephone conversation, among others) in Brazilian Portuguese, as represented in a 3.4 million word sample of the 1-billion-word Brazilian Corpus. The steps taken in the analysis, with respective findings were: (1) The corpus was tagged for selected features, using manual, automatic (the state of the art Palavras tagger) or semi-automatic procedures, the output was checked for accuracy, and corrections were made when necessary. (2) Counts were taken for each feature, which were then normalized, and standardized -- a number of variables were dropped and/or collapsed into other variables as the result of data screening; special scripts were written to count tags and display them in a user-friendly manner to allow close inspection of particular variables; (3) An initial factor analysis was run, and the number of factors in the data was determined; (4) A subsequent rotated factor analysis was conducted; (5) Factors scores were computed for each text on each factor; (6) Factors were interpreted in terms of four underlying dimensions of variation: Immediate vs elaborated reference; Reported content vs procedural focus; Involved vs informational production; Narrative vs non-narrative concerns. These represent the initial set of dimensions that account for variation in Brazilian Portuguese. (7) Each individual register was placed along each dimension so that its position with respect to the poles of the dimension was determined, as well as its relationship to the other registers in the corpus. (8) The same method had been followed for a description of written registers, which enabled a comparison with the spoken registers. This paper will show the similarities and differences between written and spoken registers in Brazilian Portuguese with respect to the dimensions found. (9) The percentage of variation that was accounted for by each dimension was calculated, and this ranged from R2 = .87 to R2 = .37, all statistically significant at $p<.000$, suggesting registers differ significantly among themselves with respect to the dimensions. The most typical texts of each dimension will be shown to illustrate the range of structural, communicative and discoursal properties featured on individual dimensions.

Research funded by Fapesp (Sao Paulo State Research Foundation) and CNPq (Brazilian National Research Council).

References

Besnier, N. (1988). The linguistic relationships of spoken and written nukulaelae registers. Language, 64, 707-736.

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

Biber, D. (Producer). (2004) Conversation text types: A multi-dimensional analysis. In lexicometrica. Jadt 2004 parcours thématique. Podcast retrieved from http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2004/pdf/JADT_000.pdf.

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

Biber, D., Finegan, E., Oostdijk, N., & de Haan, P. (1994). Intra-textual variation within medical research articles Corpus-based research into language (pp. 201-222). Amsterdam: Rodopi.

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

Crossley, S., & Louwerse, M. M. (2007). Multi-dimensional register classification using bi-grams. International Journal of Corpus Linguistics, 12(4), 453-478.

de Mönnink, I. M., Brom, N., & Oostdijk, N. H. J. (2003). Using the mf/md method for automatic text classification. In S. Granger & S. Petch Tyson (Eds.), Extending the scope of corpus based research : New applications new challenges (pp. 15-25). Amsterdam: Rodopi.

Kim, Y.-J., & Biber, D. (1994). A corpus-based analysis of register variation in korean. In D. Biber & E. Finegan (Eds.), Sociolinguistic perspectives on register (pp. 157-181). 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. Tese de doutoramento, Department of Linguistics and Modern English Language, Lancaster University, UK.

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.

Quaglio, P. (2009). Television dialogue: The sitcom friends vs. Natural conversation. Amsterdam: John Benjamins.