Abs-1

Tony Berber Sardinha, Cristina Mayer Acunzo, Marcia Veirano Pinto, Patricia Bértoli-Dutra, Renata Condi de Souza

Recent perspectives on multi-dimensional analysis

Long Colloquium (4 hours)

Title: Recent Perspectives on Multi-Dimensional Analysis

Rationale:

Multi-Dimensional Analysis (MDA) 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, MDA typically reveals a multitude of dimensions, each representing a scale of variation. MDA makes extensive use of statistical techniques, notably Factor Analysis (FA), 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 colloquium, all papers address this basic research question: what are the underlying dimensions of variation in the corpus? The basic methodology is the following: (1) The corpus is tagged for selected features, using manual, automatic or semi-automatic procedures, the output is checked for accuracy, and corrections are made if necessary; (2) Counts are taken for each feature, which are then normalized, and standardized; (3) An initial FA is run, and the number of factors in the data is established; (4) A subsequent rotated FA is conducted for the specified number of factors; (5) Factors scores are computed for each text on each factor; (6) Factors are interpreted in terms of underlying dimensions of variation.

We will present recent studies that answer this question while both complementing and pushing the boundaries of MDA, both in a synchronic and a diachronic perspective. The first one is a full register variation analysis of Brazilian Portuguese, a language not before documented in the MDA literature. Other studies each present a detailed look at individual registers found in the media, an area where a noticeable gap exists in previous research, including movies, songs and magazine articles. Another one investigates variation in student writing. A final study looks at variation in metaphor use, exploring a major feature of human language that has not received attention in the MDA literature. In addition to part-of-speech, studies in the colloquium incorporate semantic analysis, an annotation level not explored in the MDA literature, as well as tagging for linguistic metaphor.

Audience discussion: 40 minutes

Dimensions of variation in Brazilian Portuguese

Tony Berber Sardinha, Cristina Mayer Acunzo (Sao Paulo Catholic University, Brazil); Carlos Kauffmann (Folha de S.Paulo News Organization, Brazil)

This is 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 MDA study has been carried out on Portuguese. For this investigation, a 9.5 million word sample of the 1-billion-word Brazilian Corpus was chosen, comprising major spoken and written registers. It was tagged for POS using the Palavras tagger. The FA suggested a number of relevant factors, which will be presented and discussed in the presentation.

Dimensions of variation in Hollywood: the language of comedy and drama

Marcia Veirano Pinto (Sao Paulo Catholic University, Brazil)

This is a diachronic study of representative American movies based on a 350K-word corpus of 16 comedies and 16 dramas from 1940 to 2009. Selected variables were computed to match Biber's Dimension 1, to see to what extent the language of Hollywood pictures approached spontaneous dialog at the interactive end of the scale. Results show that neither comedy nor drama resembled authentic dialog, but both are close to personal letters, spontaneous speeches, and interviews. No incremental trend in dimension scores was found over time. However, a cluster analysis identified groupings of movies, which are interpreted in terms of what they might suggest for an understanding of naturalness in film language.

Dimensions of variation in British and American pop music

Patricia Bértoli-Dutra (UniToledo University, Brazil)

This is a diachronic study of UK and US pop music. The corpus is a 1.2-million word collection of over 6K lyrics recorded by 32 different artists from 1940 to 2009, including a variety of styles, from rock to punk to country. It was tagged for both part-of-speech, semantics (through a locally developed tagger), and lexical bundles (through specially designed software that mined the 1-trillion-word Google Corpus). FA indicated seven different dimensions. Close inspection of these dimensions showed how singers and bands and musical styles vary across time. Results show that the dimensions suggest a different view of styles, thus pointing to the market-driven nature of current style labels, which seem to cater for the interests of the music industry rather than reflect a concern for composition style.

Dimensions of variation in Time Magazine

Renata Condi de Souza (São Paulo Catholic University, Brazil)

This is a diachronic study of Time magazine. A 1.3 million-word corpus of texts was collected, comprising more than 300 texts, spread over a period of 70 years, from the 1930's to the present, and tagged for part of speech. The FA indicated two factors, which were interpreted in terms of the communicative properties of the texts over time. The results indicated that the dimensions were unrelated to time of publication, suggesting that the style of Time did not change gradually over time, at least with respect to the dimensions.

Dimensions of variation in learner language

Denise Delegá and Tony Berber Sardinha (São Paulo Catholic University, Brazil)

The corpus used for this study is the whole ICLE, the International Corpus of Learner Language, comprising 17 different nationalities. The texts were tagged for part of speech; frequencies for each feature were counted and normalized, and then mapped onto Biber's 1988 dimensions. The mean scores for each dimension were then compared across subcorpora, each representing a different mother tongue background. The results suggested a wide range of variation in student writing across ICLE subcorpora, especially with respect to the dimension for argumentation, suggesting that writing an argumentative essay in EFL is a task that learners around the world accomplish in different ways.

Dimensions of variation in metaphor use

Tony Berber Sardinha (São Paulo Catholic University, Brazil)

Although metaphor is a key characteristic of ordinary language, so far it has been ignored in MDA research. The aim of this paper is to report on an synchronic study of variation in metaphor use across several registers of Brazilian Portuguese. A 50K word corpus was tagged for part of speech and manually annotated at the word level for metaphor, using an identification system based on features of both MIP (Metaphor Identification Procedure) and MIV (Metaphor Identification through Vehicles). The manual markup was complemented by automatic analysis carried out by the Metaphor Candidate Identifier, a software program designed to retrieve metaphors from corpora. Each metaphorically used word was tagged a variable based on source domain, target domain, vehicle morphology, metaphor type, vehicle word class, vehicle probability level, among other features. Other variables were added, such as for morphosyntax and metaphor density. The FA indicated three factors, which were interpreted in terms of their communicative purpose, showing differences across registers with respect to metaphor use. These will be reported in the presentation.

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. (2004) Conversation text types: A multi-dimensional analysis. In Gérald Purnelle, Cédrick Fairon, and Anne Dister (eds.), Le poids des mots: Proceedings of the 7th International Conference on the Statistical Analysis of Textual Data, 15-34. Louvain: Presses universitaires de Louvain.

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.