A linguistic typology of American television programs

Tony Berber Sardinha (São Paulo Catholic University, Brazil) and
Marcia Veirano Pinto (São Paulo Federal University, Brazil)

Over the years, television programs have been classified in a large number of different ways, ranging from broad categories reflecting the general topic of the program (e.g. shows about cooking, politics, cars, etc.), the target audience (e.g. children's/teenagers'/ women's shows, etc.), the time of showing (morning/late night/Sunday morning shows, etc.), to more specific taxonomies based on the perceived genre of the show (soap opera, news, talk show, etc.) (Creeber, 2008; Frank, Becknell, & Clokey, 1971; Mittell, 2004; Rose, 1985; Wasko, 2010). Both general and specific taxonomies are in use today, in the television industry and in academia, to refer to groups of shows that share common features. However, to date no classification scheme has been developed that relied primarily on the linguistic features of the shows as a basis for the taxonomy. The goal of the current study is exactly to develop such a linguistic taxonomy of the verbal language of television programs shown in the United States. Our analysis is restricted to the spoken component of the television programs. An analysis of the visual and sound components would require a different method and would probably yield different results. Existing research on the verbal language of television from a corpus perspective has focused on comparing selected television registers among themselves or contrasting particular television programs to naturally occurring conversation (Al-Surmi, 2012; Bednarek, 2010, 2011, 2012; Csomay & Petrovic, 2012; Quaglio, 2009). No previous research has proposed a corpus-based linguistic taxonomy of the spoken language of television programs.

Our taxonomy has been developed from a multi-dimensional (MD) corpus-based perspective, using the dimensions of variation across American television registers uncovered by Berber Sardinha & Veirano Pinto (2014a; forthcoming). The MD framework is a corpus-based method introduced by Biber (Biber, 1988 et seq.; Berber Sardinha & Veirano Pinto, 2014b), whose goal is to identify the underlying parameters of variation among texts (the 'dimensions'). The dimensions are based on groupings of correlated linguistic characteristics. These groupings in turn are identified through a series of factor analyses of the normed counts of hundreds of linguistic features found across the texts (cf. Friginal & Hardy, 2014). The corpus employed for this analysis was the USTV corpus, consisting of 31 registers (programs), totaling 5.3 million words. The corpus was carefully designed so as to represent the multitude of programs presented on contemporary American television (terrestrial and cable). In addition, the size of each corpus section was calibrated so as to reflect the inherent linguistic variation among the texts, following Biber's (1993) proposal for corpus representativeness (cf. Berber Sardinha, 2014). As such, a pilot version of the corpus was collected, cleaned up, hand-checked and tagged for part of speech using the Biber tagger. The variation across the texts in each register was then assessed through a preliminary MD analysis, and extra texts were allocated to the registers that exhibited more variation. The final version of the corpus was then tagged with the Biber Tagger, and the counts of nearly 200 characteristics were taken with the Biber Tag Count program. The normed counts were analyzed factorially, thereby identifying four factors, which were interpreted as dimensions of variation, namely: 1. Oral involved vs. informational orality; 2. Reporting events; 3. Involved, stance-marked discourse, and 4. Emphatic, context-dependent discourse. The dimensions captured the majority of variation among the registers, namely 79.4% (dim. 1), 70.4% (dim. 2), 67% (dim. 3), and 51.5% (dim. 4). Each text in the corpus was scored on each of the four dimensions. The scores were obtained by adding the standardized frequencies of the features that loaded on the positive pole of each factor and by subtracting the features that loaded on the negative pole from the previous sum.

The linguistic typology was based on a cluster analysis of the dimension scores of each text, following Biber's (1989) proposal for text type identification (see also Berber Sardinha, forthcoming). In an MD text typology, texts types are '[g]roupings of text that are similar in their linguistic form' (Biber, 1989: 13). Text types are determined through cluster analysis, which:

> groups texts such that the texts within each cluster are maximally similar to each other in their exploitation of the textual dimensions, while each cluster is maximally distinct from the others. That is, those texts with the most similar dimension scores are grouped in each cluster. (Biber, 1989: 13)

A cluster analysis was performed on the dimension scores in SAS University Edition using the FASTCLUS procedure, which yielded disjoint clusters. Disjoint clusters were preferred as 'there was no theoretical reason to expect a hierarchical structure' in the text typology (Biber, 1989: 42). A challenge in cluster analysis is the determination of the optimal number of clusters in the data. In previous research of this kind, the Cubic Clustering Criterion statistic provided by the FASTCLUS procedure was used to 'provide a measure of the similarities among texts within each cluster in relation to the differences between the cluster' (Biber, 1989: 42). These heuristic devices 'reflect goodness-of-fit: the extent to which the texts within a cluster are similar, while the clusters are maximally distinguished.' (Biber & Kurjian, 2007: 120). An examination of the values of the CCC statistic seemed to indicate the presence of six clusters in the data. A provisional extraction of six clusters was then conducted. The texts in each cluster were distinguished with respect to the distance from the cluster centroid (Biber, 1989: 42). Core texts include more of the salient features on the cluster, whereas peripheral texts display fewer of the major characteristics of the cluster, which makes them 'relatively dissimilar to the central cluster characterization, but even more dissimilar to other clusters.' (Biber, 1989: 16). The clusters were interpreted qualitatively by considering how the major linguistic features of the different dimensions were used in the texts, in addition to the mean scores of the cluster on each dimension as well as the major registers included in the cluster. This linguistic profile was used to characterize the individual clusters as linguistic text types.

As mentioned, six provisional clusters have been identified, which have the following major characteristics. Cluster 1 includes texts that are extremely dense in information, narrative, stance-neutral and unmarked for emphasis and context-dependence. It is comprised mostly of news debate programs, live politics broadcasts and newscasts. Cluster 2 is in some ways similar to cluster 1, in that it comprises texts that are narrative and unmarked for emphasis and context-dependence, but the texts are not as informationally-dense and are stance-marked. The major registers in the cluster are religious programs, non-fiction series, and morning shows. Cluster 3 is unmarked for information and involvement, stance, and emphasis and context-dependence, and highly non-narrative. The texts are predominantly commercials, live sports broadcasts and infomercials. The remaining clusters are all involved in nature, being distinguished by the degree of involvement and the degree of markedness on the remaining dimensions. Cluster 4 incorporates texts that are involved, highly non-narrative, stance-neutral, and very emphatic and context-dependent. The major programs in the cluster are lifestyle shows and culinary programs. Cluster 5 corresponds to texts that are highly involved, highly narrative and highly stance-marked, and the major programs are soap operas, competition reality shows and miniseries. Finally, cluster 6 includes texts that are extremely involved, non-narrative, stance-marked and non-emphatic. The major programs are various children's programs and game shows. As mentioned, this is analysis is not final; we intend to look at ways of improving it before a final taxonomy is reached. Overall, this working typology of American television programs differs from previous taxonomies with respect to both the number of types identified and the categories determined. Furthermore, to the best of our knowledge, this is the first linguistic typology of television programs, and one of the few MD taxonomies of texts of any kind developed so far. In the paper presentation, examples of each cluster will be provided, in addition to interpretive labels of the clusters and a detailed discussion of the results.

References

Al-Surmi, M. (2012). Authenticity and TV Shows: A Multidimensional analysis perspective. *TESOL Quarterly, 46*(4), 671-694.

Bednarek, M. (2010). *The Language of Fictional Television: Drama and Identity*. London: Continuum.

Bednarek, M. (2011). The language of fictional television: A case study of the 'dramedy' Gilmore Girls. *English Text Construction, 4*(1), 54-83.

Bednarek, M. (2012). "Get us the hell out of here": Keywords and trigrams in fictional television series. *International Journal of Corpus Linguistics, 17*(1), 35-62.

Berber Sardinha, T. (2014). 25 years later: Comparing Internet and pre-Internet registers. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-Dimensional Analysis, 25 Years on: A Tribute to Douglas Biber* (pp. 81-105). Amsterdam/Philadelphia, PA: John Benjamins.

Berber Sardinha, T. (forthcoming). Text types in Brazilian Portuguese: A multidimensional perspective. *Corpora*.

Berber Sardinha, T., & Veirano Pinto, M. (2014a). *Dimensions of variation across American television registers*. Paper presented at the American Association for Corpus Linguistics Conference, Flagstaff, AZ, USA.

Berber Sardinha, T., & Veirano Pinto, M. (Eds.). (2014b). *Multi-Dimensional Analysis, 25 Years on: A Tribute to Douglas Biber*. Amsterdam/Philadelphia, PA: John Benjamins.

Berber Sardinha, T., & Veirano Pinto, M. (forthcoming). Dimensions of variation across American television registers. *International Journal of Corpus Linguistics*.

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (1989). A typology of English texts. *Linguistics, 27*, 3-43.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8*(4), 243-257.

Biber, D., & Kurjian, J. (2007). Towards a taxonomy of web registers and text types: a multi-dimensional analysis. In M. Hundt, N. Nesselhauf & C. Biewer (Eds.), *Corpus Linguistics and the Web* (pp. 109-132). Amsterdam / New York, NY: Rodopi.

Creeber, G. (Ed.). (2008). *The Television Genre Book*. London: British Film Institute.

Csomay, E., & Petrovic, M. (2012). "Yes, your honor!": A corpus-based study of technical vocabulary in discipline-related movies and TV shows. *System, 40*(2), 305-315.

Frank, R. E., Becknell, J. C., & Clokey, J. D. (1971). Television program types. *Journal of Marketing Research, 8*, 204-211.

Friginal, E., & Hardy, J. A. (2014). Conducting Multi-Dimensional analysis using SPSS. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-Dimensional Analysis, 25 Years* *on: A Tribute to Douglas Biber* (pp. 298-316). Amsterdam/Philadelphia, PA: John Benjamins.

Mittell, J. (2004). *Genre and Television: From Cop Shows to Cartoons in American Culture*. London/New York: Routledge.

Quaglio, P. (2009). *Television Dialogue: The Sitcom Friends vs. Natural Conversation*. Amsterdam; Philadelphia, PA: John Benjamins.

Rose, B. G. (Ed.). (1985). *TV Genres: A Handbook and Reference Guide*. Westport, Conn.: Greenwood Press.

Wasko, J. (Ed.). (2010). *A Companion to Televison*. London: Wiley-Blackwell.