Clusters and Correspondences. A comparison of two exploratory statistical techniques for semantic description

Introduction

Corpus-Driven quantitative techniques for language description have witnessed important success in recent years. In semantic research, the main of this drive has been in disambiguation, whether on the syntagmatic or paradigmatic plane. Many have now taken the next step, seeking to employ such techniques for the description of lexical semantic structure per se. This study examines two exploratory multivariate statistical techniques, namely Multiple Correspondence Analysis (MCA) and Hierarchical Cluster Analysis (HCA), and considers the strengths and weaknesses of both approaches for the description of lexical semantic variation.

This study is informed by the usage-based approach of Cognitive Linguistics, represented by Geeraerts & al. (1994), Gries (2003), Geeraerts (2006), Tummers & al. (2005), Gries & Stefanowitsch (2006), Grondelaers & al. (2007), and Heylen & al. (in press). Within this field, both exploratory and confirmatory statistical techniques enjoy wide currency. Four of the main techniques include Cluster Analysis (Divjak 2006, Divjak & Gries 2006, Gries 2006), Correspondence Analysis (Arppe 2006, Glynn in press, Gries & David forthc.) for exploratory research and Logistic Regression Analysis (Heylen 2005, Tummers & al. 2005) and Linear Discriminant Analysis (Gries 2001, Wullf 2003) for hypothesis testing.

Case Study

Using a large non-commercial corpus, built from on-line personal diaries and specified for the regional difference (American vs. British English), the study examines the semantic structure of the lexeme annoy, semasiologically and onomasiologically in comparison with two parasynonymous words; hassle and bother. Approximately 500 occurrences are manually annotated for 20 formal, semantic, and extra-linguistic variables. An important challenge for corpus linguistics is semantic description. In order to maximise objectivity of the semantic annotation, special attention is given to the Frame Semantic actor types and their relations. This method has been shown to provide indirect indices of semantic structure (Glynn & al. forthc). Despite the regional variation, the corpus is quite homogenous in terms of register and theme. However, this allows us to focus on the dimension of dialect variation, relative to the variables of morpho-syntax and Frame Semantic argument structure.

At an exploratory level, both MCA and HCA have important strengths and weaknesses. One important difference between the two techniques is that Cluster Analysis is primarily designed to present its results in the form of dendograms where Correspondence Analysis relies on scatter plots. The dendograms of HCA offer clear representations of both the grouping of features and the relative degree of correlation between those features. The trees represent relations and the shorter the distance between the node and the branch, the higher the degree of correlation. The principle shortcoming of this representation it that gives the false impression that all the data fall into groups, where in fact this may not be the case. Figure 1, offers an example of the
dendogram of HCA. The table clusters the lexmes – *annoy* – *bother* - *hassle*, with by variables of dialect – grammatical construction, grammatical class, cause and affect of the event.

**Figure 1. HCA annoy-hassle-bother**

The scatter plots of Correspondence Analysis, although at times difficult to interpret, offer a much more 'analogue' representation of correlation. In such plots, the spatial dispersion and relative proximity of data points represent degrees of correlation. The interpretation of the plot is then much more approximative than the dendogram. Figure 2. offers an example of the visualisation of an MCA analysis for the same variables that are treated above.

**Figure 2. MCA annoy-hassle-bother**
The results between the two techniques here seem somewhat in contrast. In order to better appreciate the differences. A simpler dataset is analysed. Looking at just hassle, we get a more comparable result. Agent and Patient Types, Agent and Patient Person (1st, 2nd, 3rd) as well as Agent-Patient relations differ significantly in the use of annoy between the two dialects. Moreover, these differences are mirrored by purely semantic variables. The combination of the Frame Semantic, formal and traditional semantic variables show that annoy possesses a more emphatic and 'anger' related meaning in American than in British English, where its use is lighter and less likely to be used for situations to describe serious malcontent. The two methods, to varying degrees, confer on these results.

These results are then compared to those of a stepwise logistic regression analysis using the dialect distinction as a response variable. The regression analysis clearly shows that MCA, despite the complexity of the scatter plots, better captures the relative associations revealed in the data. This is most likely due to the need to conflate variables in HCA architecture which in turn may cause low frequency cells. This may explain why the results of the HCA seem to “lump together” less frequent correlations. Although the regression model reveals that the results of the MCA are more informative, the correlation of certain data points is erroneously represented. At one point, a rarely occurring, but crucial, feature is shown to be associated with one dialect, where in fact this merely results from a superficial effect of the two dimensional representation of multidimensional space; the feature in question being “drawn towards” the dialect variable because of its association with another non-relevant data point.

In conclusion, the comparable results of both methods demonstrate their usefulness as exploratory techniques. However, both HCA and MCA can be unreliable when faced with complex multivariate data. In light of this, their results warrant confirmatory analysis. Nevertheless, the contrast in the results of the complicated analysis across the three lexemes, suggest that MCA is better suited to truly multivariate exploratory research. One possible advance for these techniques lies in integrating bootstrap resampling and expectation maximisation. Implementing such algorithms may resolve some of the concerns that these exploratory methods face.

References


