Semantic Fields And Polysemy: Essay, Research Paper

Semantic Fields and Polysemy: A Correspondence Analysis Approach

Semantic Fields

Studying semantic fields or literary themes in texts quickly confronts the researcher with a paradox. A computer string search will produce a list of the frequencies of words potentially related to the semantic field. But polysemy–the fact that many words have multiple significations–means that there is an un-measured difference between the potential and the real allusions.

The semantic field of “solitude” or “loneliness” is a case in point. Important from a sociological and psychological perspective as an indication of imperfect adaptation to one’s milieu, it is also a frequently occurring literary theme. In French, the problem of polysemy is acute because “seul” means both “alone” and “only”, so it is difficult to imagine that the expression “ma seule cravate (my only tie)” having much to do with solitude. Similarly, the verb “abandonner (to abandon)” is indistinguishable by computer from “s’abandonner (to let oneself go).

The Problem of Disambiguation

The only practical response to such difficulties seems to be disambiguation by human informants. But one can legitimately be concerned about the reliability of such a process. The consistency of results from one informant to another when examining the same data would seem to be a reasonable touch-stone for the reliability of the disambiguation process. It seems appropriate to expect some personal variability among informants, so the choice of statistical methods to be used requires some care.

An intra-class correlation coefficient similar to Cohen’s kappa has been chosen as a measure of agreement among the individual choices made by each informant on each word potentially evoking solitude. Correspondence analysis has been used in order to provide a visual representation of how the data interact.

Data

Nine 20th century novels written in the first person were chosen for this analysis: Bernanos, Journal d’un cur de campagne, Camus, L’ tranger and La Chute, C line, Voyage au bout de la nuit, Gide, L’Immoraliste and La Porte troite, Mauriac, Le Noeud de Vip res, Proust, La Fugitive, and Sartre, La Naus e. The size of these novels ranges between 31,272 and 192,559 words.

Some seventy words (in the sense of lemmas) related to the concept of solitude in French thesauri can be identified. They reduce to thirty strings of the type “seul\*”. These strings were used to search the texts for words related to solitude in the ARTFL database. Results ranged between 473 and 73 occurrences, and it must be recalled that these numbers relate to potential evocations of solitude only.

The words found by the ARTFL search engine, centred in 60 characters of context were downloaded, and given to a team of informants with minimal instructions: choose the words evoking human solitude from a reading of the context, and go back to the ARTFL database for more context in doubtful cases. Eight informants were used: two French literature professors, two French literature graduate students whose native language was French, two French literature graduate students whose native language was English, and two high school students who had taken immersion French. They provided results ranging between 122 and 11.

Results

Analysis of the frequencies provided by the informants as representing the true allusions to solitude in the nine texts demonstrates that the frequencies provided are not a linear function of the number of potential allusions to solitude. Pearson’s correlation coefficient, and the chi- squared contingency table test would seem like appropriate analytic tools, but they have disadvantages. The examination of either the correlation coefficient table or the individual chi-squared values making up the chi-squared statistic for trends or tendencies is rendered quite difficult by the volume of the data (nine novels and eight informants).

These results have the disadvantage of applying to aggregate data, collapsing into a single total what may well have been different individual choices. The intraclass correlation coefficient is a measure of agreement among individual choices. Although it has been modified using Cronbach’s alpha to take into account that a majority opinion is been used as a standard, it is similar to Cohen’s kappa, and like the latter measure, a value of 0.55 or greater can be taken as indicative of good agreement among individual choices (May). The score for the seven informants for whom individual scoring data were available was 0.55125. Application of the Cohen’s kappa routine in the JMP-IN software on a pair-wise basis produces low readings for Student 2 (data for Student 1 were not available). In the light of these inconsistent results a means of visualising what is going on in the data is desirable, and correspondence analysis was chosen for this.

The correspondence analysis technique (Benz cri, Greenacre) is mathematically complex, but widely available. Essentially, it provides to the user representation of the variability of the data by projecting onto a two-dimensional space the contributions of both the rows and the columns of the chi-squared contingency table to the chi-squared statistic in such a way that the further two points are apart the greater their difference, and the closer they are the greater the similarity of the distributions they represent. Figure 1 illustrates the relationship of the frequencies for solitude by all eight informants, as well as the raw frequencies of the data on which they worked. The position of the raw frequencies at the extreme left of the map and of Student 2 in the upper right quadrant clearly identify them as outliers. The rest of the informants are clustered in the lower right quadrant.

Figure 2 shows the same frequencies after the raw frequencies and Student 2 have been removed, as well as Student 1 whose consistency could not be measured on the basis of individual choices. The structure of the data is manifest. There is variation, owing to differences in judgment. The texts group well on the left side of the map with Sartre and Camus, existentialist writers occupying this space. Proust, Gide and Mauriac, all bourgeois authors, are at the lower right. C line and Bernanos both right-wing critics of society are in the upper right quadrant.

Most important, the informants do not cluster according to level of education or linguistic background, as shown by the distance between p1 and p2, as well as between s3 and s4. In short, the data present no clear pattern on the basis of background or level of education, and variation can be reasonably ascribed to differences in personal interpretation.

Conclusion

Most of us study language and literature by computer because we have deep-seated reservations about techniques that rely simply on the impressions of the researcher. Many of us, particularly in literature, are reluctant to hand over to student assistants the job of doing preliminary analysis of material on which we will subsequently base our interpretations. Many of us would prefer to do the work ourselves rather than rely on the opinion of others. Linguists, on the other hand, have long used native speaker informants.

The results reported here illustrate the usefulness of correspondence analysis for interpreting complex data. They also suggest that a person with native-speaker ability in a language, even an originally English-speaking graduate student in French, will produce about the same results as a professor of French literature. It would seem then that the use of informants for studying semantic fields, or literary themes, is a justifiable enterprise from the statistical perspective.