How to determine the meaning and use of (causal) connectives in (large) corpora : from hand-based to automatic analyses

Liesbeth Degand, Yves Bestgen

How to determine the meaning and use of (causal) connectives in (large) corpora: from hand-based to automatic analyses

Liesbeth Degand & Yves Bestgen

FNRS/Université catholique de Louvain, Louvain-la-Neuve, Belgique

degand@lige.ucl.ac.be, yves.bestgen@psp.ucl.ac.be

The main objective of this contribution is to present a number of methodological problems arising in the collection and interpretation of data from (large) corpora. In particular, we want to show how large corpora can be used to test hypotheses concerning linguistic factors determining the meaning and use of connectives, focusing on a number of methodological issues.

Corpus-based approaches to connectives are not new, but classically they consist of either fully analysed but relatively small corpora, or of large corpora of which a random set is analysed. The reason for this quantitative restriction is clear: the data-analyses are completely hand-based. While these empirical studies are useful from a qualitative point of view, they all suffer from the same quantitative drawback, namely the relatively small number of data (rarely more than 100 occurrences are analysed, mostly only 50). In addition, most of these analyses are still too analyst-dependent, making generalizations and replications difficult.

In ongoing research we propose to change this situation by handling exhaustively large corpora (with hundreds and even thousands of occurrences of the same linguistic phenomenon) and by implementing the analysis procedures to make them analyst-independent (Bestgen, Degand & Spooren 2003, submitted). This is done by exploiting prior linguistic results with their level of detail and sophistication but adding a much larger scale of analysis. In short, our general purpose is to enrich qualitative linguistic analyses with massive quantitative data. To achieve this goal, we make use of two components: a series of linguistic hypotheses to test, derived from the literature on connectives, and a number of NLP techniques to analyse the data, in our case Latent Semantic Analysis (Landauer, Folz & Laham, 1998) and Thematic Text Analysis (Popping, 2000).

The combination of these two techniques has proven to be successful. That is, we succeeded in confirming a number of linguistic hypotheses resulting from prior small-scale corpus analyses. According to these results, the meaning and use of (Dutch causal) connectives in newspaper corpora is determined by a series of factors including the level of subjectivity of the environing context, the flow of discourse (topicality), and the degree of polyphony of the connective. It is the topic of ongoing research to refine a number of these results and to apply the method to the use and meaning of French causal connectives in different text genres. Part of these results will be presented at the workshop.

These first successful results should not hide that the method is not free of methodological flaws. The main issue we would like to raise concerns the selection and interpretation of the data to analyse. This problem arises in hand-based analyses first, if the procedure is implemented, it arises in automatic analyses too. With respect to the analysis of (causal) connectives the specific questions that could arise are the following: (i) should all the observed data be included in the analyses (what about the "non-causal" use of "causal" connectives?), (ii) which problems arise in (hand-)coding the data (how many different causal relations?, what about analyst-dependent interpretations? (the problem of low interrater agreement), what is the minimal sample of data to analyse?, …), (iii) what is the relationship

between fine-grained linguistic analyses and LSA-base statistic analyses? (with the particular problem of the determination of the environing context which can be intuitive in hand-based analyses, but must be implemented in automatic analyses), and finally (iv) what contributions can the two methods (hand-base and automatic) make to one another.

References

Bestgen, Yves, Degand, Liesbeth & Spooren, Wilbert (2003). On the use of automatic techniques to determine the semantics of connectives in large newspaper corpora: an exploratory study. Lagerwerf Luuk, Spooren Wilbert, Degand Liesbeth (Eds). *Determination of Information and Tenor in Texts: Multidiscplinary Approaches to Discourse 2003,* Stichting Neerlandistiek VU Amsterdam & Nodus Publikationen Münster, 179-188.

Bestgen, Yves, Degand, Liesbeth & Spooren, Wilbert (submitted). Towards automatic determination of the semantics of connectives in large newspaper corpora. *Discourse Processes*.

Landauer, T.K., Foltz, P.W., & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes*, *25* (2, 3), 259-284.

Popping, R. (2000). *Computer-assisted Text Analysis*. London: Sage.