A profile-based calculation of region and register variation: the synchronic and diachronic status of the national variants of Dutch

Dirk Speelman, Stefan Grondelaers & Dirk Geeraerts
Department of Linguistics, Blijde-Inkomststraat 21, B-3000 Leuven, Belgium
dirk.speelman@arts.kuleuven.ac.be

Over the last decade, *the Leuven Research Unit of Quantitative Lexicology and Variational Linguistics* has developed a corpus-based method for the investigation of region and register variation in and between the national variants of Dutch, viz. Belgian Dutch and Netherlandic Dutch. The basic characteristics of the methodology are the following. Geeraerts, Grondelaers & Speelman (1999) introduces "linguistic uniformity" as an operational measure to determine the lexical distance between language varieties. Uniformity is calculated by quantifying the overlap between onomasiological profiles. (The onomasiological profile of a particular concept in a particular language variety is the set of synonymous alternative terms used to name that concept in that language variety, differentiated by the relative frequency of the alternative terms in a corpus.) These uniformity-based calculations have in the past been used in top-down analyses of convergence and stratification. More recently, they are incorporated in a broader framework of multivariate statistical analysis to achieve a bottom-up classification of language varieties.

In this talk we want to present **the technical basis of profile based calculations**. After explaining the actual calculations, we illustrate the merits of the approach. By taking all alternative designations of a phenomenon (and their frequencies) into account, we can (i) put the statistical importance of isolated variables into perspective, (ii) avoid the blurring effect of polysemy and semantic ambiguity of isolated variables, and (iii) avoid confusion between the region & register differences that interest us, and thematic, situational or medium-specific biases in the corpus.

From the list of practical issues that need to be tackled before profile-based uniformity calculations can safely be extrapolated to any kind of lexical and non-lexical indicators (an extension to function words can be found in Grondelaers e.a. 2001), we zoom in on the topic of *variable selection*. We show how this selection is facilitated by the automated generation of exhaustive lists of 'stable' markers of specific registers and regions (which are derived from all possible two-by-two comparisons of the word or n-gram frequency lists of the corpora that represent the registers and regions at issue - in this preparational step statistical significance is established with the method explained in Dunning 93.

References

Dunning T. 1993. ‘Accurate Methods for the Statistics of Surprise and Coincidence’. *Computational Linguistics* 19(1): 61-74.

Geeraerts D., Grondelaers S., Speelman D. 1999. *Convergentie en divergentie in de Nederlandse woordenschat. Een onderzoek naar kleding- en voetbaltermen*. Amsterdam: Meerstensinstituut.

Grondelaers S., van Aken H. Speelman D., Geeraerts D. 2001. ‘Inhoudswoorden en prepositie als standaardiseringsindicatoren. De diachrone en synchrone status van het Belgische Nederlands’. Nederlandse Taalkunde 6: 179-202.