The sem·metrix Project: Towards the Large-Scale Measurement of Lexical Variation

Yves Peirsman[1] and Kris Heylen

Abstract

In this paper we present the methodology and first results of the sem·metrix project, which aims to develop corpus-based tools for the large-scale measurement of lexical variation. The project takes as its starting point the profile-based study of lexical variation developed by Geeraerts, Grondelaers and Speelman (1999). Within this approach, a profile is defined as a set of near-synonyms that express a certain concept, together with their relative frequencies in a corpus. On the basis of representative corpora, these profiles can be used to measure lexical variation between language varieties with respect to the profile concept.

Geeraerts, Grondelaers and Speelman (1999) successfully investigated the lexical variation between Belgian and Netherlandic Dutch in the semantic fields of clothing and football. However, their original approach proved difficult to extend because of the time-consuming manual definition of profiles. The sem·metrix project tries to overcome this problem by using a number of computational-linguistic algorithms.

In the first part of our talk, we will present the methodology behind the project. To generate profiles, the sem·metrix project relies on established computational-linguistic methods that use distributional similarity to identify semantically related words. These are then clustered into sets of near-synonyms.

In the second part of the presentation, we will illustrate this automated profile-based methodology with a case study. Following Geeraerts, Grondelaers and Speelman (1999), we will look at lexical variation in the naming of football concepts in Belgian and Netherlandic Dutch and investigate whether the results of the manual analysis can be replicated with the automatic method.

References

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

[1] QLVL, University of Leuven
e-mail: yves.peirsman@arts.kuleuven.be