A distributional semantic approach to the identification of stages in constructional productivity change

Florent Perek, University of Birmingham

Much work in historical linguistics involves identifying stages of language change: periods of relative stability in the recorded usage of some construction and times of shift. In recent work, Gries & Hilpert (2008) proposed a quantitative usage-based approach to this issue, called variability-based neighbour clustering (VNC). VNC is a customized version of agglomerative clustering that consists in aggregating adjacent periods that are closely similar in terms of some quantitative criteria measured on the relevant construction in diachronic data. The output of VNC is a partition of the time scale into periods that are maximally coherent with respect to the relevant criteria.

Most applications of VNC so far have taken as their basis purely quantitative information: token frequency, type frequency and other measures derived from them, or the frequency distribution of lexemes occurring in one of its slots. However, information of this kind does not directly capture semantic dimensions of change, such as whether the construction is used with different semantic classes of lexical items.

This paper presents an extension of VNC that addresses these limitations by drawing on a distributional semantic model as a proxy to word meaning. Drawing on the observation that words occurring in similar contexts tend to have similar meanings, distributional semantic representations approximate the meaning of a word by recording its co-occurrence with other words in a vast corpus (Turney & Pantel, 2010). The present approach consists in adding the distributional representations of all words occurring in a construction at different points in time, and using the resulting combinations as input to VNC. The method is illustrated by a case study on the recent history of the *way*-construction, showing how VNC identifies successive periods of productivity during which the construction has gradually attracted more abstract classes of verbs.

Turney, P. & Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research* 37, 141-188.
Gries, S. & Hilpert, M. (2008). The Identification of Stages in Diachronic Data: Variability-based Neighbor Clustering. *Corpora* 3: 59–81.