Finding Groups in Chronologically-Ordered Corpus Data: Variance-Based Neighbor Clustering

Stefan Th. Gries[1]

Abstract

Much corpus-linguistic research is concerned with the development of particular parameters over time. For example, in L1/L2 acquisition/learning, the syntactic development of a child/learner is approximated on the basis of how mean lengths of utterances (MLU), *t*-unit-based measures, or IPSyn values change over time (cf. Shirai and Andersen 1995 or Ortega 2003). Similarly, in historical linguistics, an expression's degree of grammaticalization is often approximated via the percentages of the expression's use as a lexical or grammatical element change over time; cf. Svensson (2000).

Most such studies aim at representing this variation in terms of stages. The probably best known example is that of Brown's (1973) MLU stages, which underlie much work on L1 acquisition, but cf., say, Hilpert (2006) for a diachronic example. However, so far no broadly applicable yet principled method to do this has been developed. In language acquisition, Brown's cut-off points are essentially arbitrary; elsewhere, the data are just split up into *n* equally-sized parts, where *equally-sized* variously refers to amounts of time or numbers of items.

I will introduce a completely data-driven method, variance-based neighbor clustering (VNC), that takes as input chronologically-ordered corpus data and solves this problem. It is similar to standard clustering approaches because clustering is performed objectively using quantitative information and represented graphically in the form of dendrograms. VNC differs from standard approaches because it only clusters neighboring data points, thus preserving the data points' temporal sequence. I will discuss the advantages of this methods on the basis of results from two submitted case studies, one based on the tense-aspect acquisition from the Stoll corpus of Russian L1 acquisition, the other based on data regarding infinitives following *shall* from the Penn Parsed Corpora of Historical English

References

Brown, Roger. 1973. *A first language: the early stages*. Cambridge, MA: Harvard University Press.

Hilpert, Martin. 2006. Distinctive collexeme analysis and diachrony. *Corpus Linguistics and Linguistic Theory* 2.2.

Ortega, Lourdes. 2003. Syntactic complexity measures and their relationship to L2 proficiency: a research synthesis of college-level L2 writing. *Applied Linguistics* 24.4:492–518.

Shirai, Yasuhiro and Roger W. Andersen. 1995. The acquisition of tense-aspect morphology: a prototype account. *Language* 71.4:743–62.

[1] *e-mail*: stgries@linguistics.ucsb.edu

Svensson, Patrik. 2000. Grammaticalization of partitive constructions. Seminar given at the Department of English, Uppsala University.