A statistical grammar as empirical resource for inducing lexical semantic phenomena

Sabine Schulte im Walde & Heike Zinsmeister
Universität des Saarlandes, Universität Tübingen
schulte@coli.uni-sb.de, heike.zinsmeister@uni-tuebingen.de

Our talk presents the grammar modelling framework of *Head-Lexicalised Probabilistic Context-Free Grammars (HeadLex-PCFGs)*, cf. Charniak (1995) and Carroll and Rooth (1998), as an empirical resource for inducing quantitative lexical properties at the syntax-semantics interface. As the core of a HeadLex-PCFG, a context-free grammar is developed, with head-marking on the children. The parameters of the probabilistic version of the context-free grammar – both for the unlexicalised PCFG, a lexicalisation bootstrapping, and the lexicalised HeadLex-PCFG – are estimated in an unsupervised training procedure, using the Expectation-Maximization algorithm (Baum, 1972). The trained grammar model provides lexicalised rules and syntax-semantics head-head co-occurrences; the lexicalised parameters enable the induction of semantic phenomena.

Properties of the German HeadLex-PCFG: The German context-free grammar was developed with specific attention towards verb subcategorisation and therefore shows a verb rule bias: out of 35,821 grammar rules, 94% refer to verb phrase rules; the remaining 6% cover noun, adjective, adverbial, and prepositional phrases. Most grammar rules are binary rules, i.e. rules with only two children, which ensures a step-by-step analysis of verb phrase saturation. Rules modelling closed-class items, such as prepositional phrases, incorporate multiple features into the grammar categories; e.g. the PP categories are tripartite, including the category name, the preposition itself, and the case of the subcategorised phrase.

The German grammar concentrates on mass phenomena; i.e. in favour of covering linguistic properties of a large amount of lexical items, selected irregularities were disregarded. For example, we did not model the subcategorisation of genitive or adjectival arguments, which is required for a specific subset of verbs only. In addition, the rules deliberately contain a considerable degree of ambiguity, in order to train lexeme-specific preferences; for example, prepositional phrases can apply to verb phrase rules both as arguments or as adjuncts, which learns the verb-specific distinction.

Inducing lexical semantic phenomena: We demonstrate the usage of the grammar model for inducing various lexical properties at the syntax-semantics interface. For example, Schulte im Walde (PhD-Thesis, 2003) defined subcategorisation properties for verbs with respect to frame types and selectional preferences, and used them to induce semantic verb classes, and Schulte im Walde (COLING, 2004) showed that the subcategorisation information is also available for the specific case of particle verbs; Zinsmeister (Corpus Linguistics, 2003) identified predicatively used adverbs, an idiosyncratic property of the lexicalised items, and classified the adverbs into semantic functions. Zinsmeister and Heid (KONVENS, 2004) induced verb-noun collocations from the grammar model, and applied them to distinguish compositional vs. idiomatic noun compounds. Additional potential for semantic properties includes e.g. word order preferences with respect to verbs and clause types, the induction of semantic adjective classes as based on adjectival modifiers and head-head combinations, the distinction of PP arguments and adjuncts and inducing semantic PP functions, and the induction of adverbial classes on basis of their clause modification.

Discussion of design properties: The design and the exploitation of the German HeadLex-PCFG are quite idiosyncratic, as compared to the induction of lexical properties from manually annotated treebanks, or from (Viterbi) parses of a (statistical) grammar. Specifically, we would like to discuss the general HeadLex-PCFG framework as basis for lexical acquisition, the design criteria of mass vs. detailed phenomena description, the deep vs. flat analysis structures, the advantages and disadvantages of introducing ambiguity into grammar rules, and the development and training effort.