

# Functional features in data-driven parsing

The separation of functional structure from constituent structure is motivated largely by cross-linguistic variation in level of configurationality: languages differ in the extent to which grammatical functions may be equated with a specific structural position. F-structure constraints capture generalizations regarding grammatical functions regardless of their c-structure realization. In functional-typological Optimality Theory (Aissen, 2003; Bresnan and Aissen, 2002) constraints targeting grammatical functions have been centered around a notion of prominence and harmony, which have been shown to capture both categorical generalizations, as well as frequency effects observed in a range of languages (Bresnan et al., 2001). The idea that grammars are inherently probabilistic in nature has been motivated by empirical evidence observed as frequency effects in linguistic studies ranging from computational, psycholinguistic, typological to more theoretical (Bresnan, 2006; Manning, 2003; Bod, 1998). In computational linguistics, data-driven, statistical methods show impressive results for a range of NLP tasks, including syntactic parsing. There exists an expressed interest in a deeper understanding of the results obtained using data-driven methods and how these relate to generalizations from more theoretically oriented work.

The Scandinavian languages are characterized by variation in the positioning of arguments and adverbials, in combination with rigid verb placement. These properties have led to the proposal of a considerably flatter c-structure than previously assumed for these languages (Börjars et al., 2003). The examples in (1)-(2) illustrates word order variation which departs from the canonical SVO ordering of arguments.

- (1) *Samma erfarenhet gjorde engelsmännen*  
same experience made englishmen-DEF  
'The same experience, the Englishmen had'
- (2) *Därefter betalar patienten avgift med 10 kronor*  
thereafter pays patient-DEF fee with 10 krona-PL  
'Thereafter, the patient pays a fee of 10 kronas'

In (1), the direct object occupies sentence-initial position, and in (2) an adverbial resides initially, forcing the subjects and other arguments to post-verbal position(s).

This paper will present experiments in data-driven dependency parsing of Swedish. The focus will be on the analysis of syntactic arguments and, in particular, on *argument differentiation*: the process by which functional arguments are distinguished along one or more linguistic dimensions. In a data-driven parser, parsing is per definition guided by frequencies in language and there is no explicit grammar. This allows us to make as few assumptions as possible with respect to formulations of constraints on arguments, as well as their interaction. Due to the variation identified above, we do not want to commit to a strictly structural definition of argument status. Rather, a view of grammatical functions as primitive notions, separated from surface linguistic properties, enables investigations also into mismatches between levels of linguistic analysis.

We use the freely available MaltParser,<sup>1</sup> which is a language-independent system for data-driven dependency parsing, which is based on a deterministic parsing strategy, in combination with treebank-induced classifiers for predicting parse transitions (Nivre, 2006). It allows for explicit formulation of features employed during parsing by means of a feature model. The parser is trained on the Swedish treebank Talbanken05, which consists of around 200,000 running tokens with syntactic analysis in dependency format Nivre et al. (2006). In dependency analysis, exemplified by Figure 1, functional argument structure is separated from structural positioning and formulated as dependency relations between lexical elements. Structural assumptions are furthermore stripped down to the minimal relation between a head and its dependent (Mel'čuk, 1988).

---

<sup>1</sup><http://w3.msi.vxu.se/users/nivre/research/MaltParser.html>

| Linguistic feature | Treebank feature    |
|--------------------|---------------------|
| animacy            | person reference    |
| definiteness       | morph. definiteness |
| referentiality     | pronoun type        |
| √2/finiteness      | tense               |
| nominal, verbal    | part-of-speech      |
| case               | morph. case         |

Figure 1: Linguistic features and their experimental counterparts.

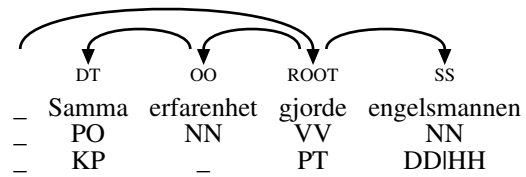


Figure 2: Example (1) from Talbanken05 with dependency annotation.

We formulate a set of linguistically motivated features which capture important aspects of the realization of grammatical functions in Swedish, see Figure 1. Features expressing the animacy, definiteness and referentiality of an argument highlight its semantic, thematic and cognitive prominence (de Swart et al., 2007; Gundel et al., 1993). Features expressing minimal structural assumptions with respect to Swedish include finiteness (Holmberg and Platzack, 1995) and features distinguishing nominal and verbal categories. These are largely lexical features, expressing local and inherent properties of lexical elements.

In a set of controlled experiments, we investigate the effect of the features in Figure 1 on the assignment of grammatical functions in Swedish. We will present an in-depth error analysis which relates the errors performed by the parser to properties of the linguistic features supplied to the parser, as well as ambiguities caused by variation in word order and lack of morphological marking. The results show clear improvements both in the performance of the parser for specific grammatical functions, such as subject, object and indirect object, as well as in overall parse performance ( $p < .0001$ ). Moreover, the effect of the various features employed during parsing correspond to a ranking, ranging from more or less categorical constraints on the realization of grammatical functions, such as finiteness, to “softer” constraints expressing preferences in function-structure mapping, such as ordering constraints (SUBJ > OBJ), as well as preferences for animate subjects and inanimate objects.

## References

- Judith Aissen. Differential Object Marking: Iconicity vs. economy. *Natural Language and Linguistic Theory*, 21:435–483, 2003.
- Rens Bod. *Beyond Grammar*. CSLI Publications. University of Chicago Press, 1998.
- Kersti Börjars, Elisabet Engdahl, and Maia Andréasson. Subject and object positions in Swedish. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the LFG03 Conference*, 2003.
- Joan Bresnan. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Sam Featherston and Wolfgang Sternefeld, editors, *Roots: Linguistics in search of its evidential base*, Studies in Generative Grammar. Mouton de Gruyter, 2006.
- Joan Bresnan and Judith Aissen. Optimality and functionality: Objections and refutations. *Natural Language and Linguistic Theory*, 20(1): 81–95, 2002.
- Joan Bresnan, Shipra Dingare, and Christopher D. Manning. Soft constraints mirror hard constraints: Voice and person in English and Lummi. In *Proceedings of the LFG '01 Conference*. CSLI Publications, 2001.
- Peter de Swart, Monique Lamers, and Sander Lestrade. Animacy, argument structure and argument encoding: Introduction to the special issue on animacy. *Lingua*, 2007.
- Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. Cognitive status and the form of referring expressions. *Language*, 69(2):274–307, 1993.
- Anders Holmberg and Christer Platzack. *The role of inflection in Scandinavian Syntax*. Oxford University Press, New York/Oxford, 1995.
- Christopher D. Manning. Probabilistic syntax. In Rens Bod, Jennifer Hay, and Stefanie Jannedy, editors, *Probabilistic Linguistics*, pages 289–341. MIT Press, 2003.
- Igor Mel'čuk. *Dependency Syntax: Theory and Practice*. State University of New York Press, 1988.
- Joakim Nivre. *Inductive Dependency Parsing*. Springer, 2006.
- Joakim Nivre, Jens Nilsson, and Johan Hall. Talbanken05: A Swedish treebank with phrase structure and dependency annotation. In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006)*, Genoa, Italy, May 24-26 2006.