Issue of Valency in Prague Dependency Treebank: Creating Valency Lexicon of Verbs

Markéta Lopatková
Center for Computational Linguistics, MFF UK, Prague
lopatkova@ckl.mff.cuni.cz

The Prague Dependency Treebank (PDT) meets the wide-spread aspirations of building corpora with rich annotation schemes. As various valency behavior of verbs cannot be described by general rules, the valency lexicon belongs to the basic resources for all rule-based task of NLP.

In principal, there are two general approaches to the description of valency – primarily syntactically-based approach (e.g. PropBank, Levin classes) and primarily semantically-based approach (e.g. FrameNet, LCS Database). The PDT, based on Functional Generative Description of Czech (Sgall et al., 1986) and its theory of valency (Panevová, 1994), has adopted a 'middle course': syntactic criteria are used for the identification of Actor and Patient, Actor is the first inner participant, the second is always Patient. Other inner participants (Addressee, Origin and Effect) as well as free modifications (about 45 for verbs) are detected in accordance with semantic considerations.

The electronic lexicon being created offers complex information on each lexical entry, verb lexeme, which includes information on particular valency frames (corresponding to their meanings) as well as information specifying elements of these frames (functors, i.e. valency relation between a verb and its complementation, possible morphemic realizations and its obligatoriness). Also additional information useful for NLP, namely reciprocity, type of control and possible diatheses, aspectual counterparts and links to Czech WordNet are introduced. For the purposes of NLP it is important to describe particular verbs in all their meanings.

An extensive application of theoretical principles of valency points out the necessity of further refinement of the theory (esp. quasi-valency and typical complementations creating enriched valency frames are presented).

At present the valency lexicon contains 1018 verbs with 2975 valency frames (i.e. 2,9 frames per verb) – the 'coverage' of the lexicon is about 85% on the verbs in running text from Czech National Corpus.