Enriching WordNet with Derivational Subnets

Karel Pala and Radek Sedláček

Faculty of Informatics, Masaryk University,
Botanicka 68a, 60200 Brno, Czech Republic
{pala, rsedlac}@fi.muni.cz

Abstract. In this paper, we deal with the derivational (word formation) relations as they are handled by the Czech morphological module Ajka. First, we show that they represent empirically well-based semantic relations forming small semantic networks, and then we solve the problem how to integrate them into lexical database such as (Czech) WordNet. In this respect we examine the relation between the derivational relations and semantic roles (deep cases) defined as Internal Language Relations in EuroWordNet. An attempt is made to match up the inventory of the semantic roles in EWN with the derivational (semantic) relations. We also use a tool called SAFT that can process a raw (corpus) text in such a way that it uses module Ajka to find links relating the WordNet senses to the noun and verbal lemmata obtained from the raw (corpus) text. This technique allows us to enrich Czech WordNet with the derivational subnets and represent them in a XML format. The result is a new kind of the semantic network, which consists of two layers, upper and lower. The result is a more **powerful** and efficient resource for applications like tools for WSD, web searching or information extraction.

1 Derivational Relations as Semantic Networks

For computer processing highly inflected language like Czech it is necessary to have a high quality morphological module that can perform lemmatization of a given word form and yield all the grammatical categories that are carried by the word form. Such a tool for Czech is a morphological analyzer and generator called Ajka developed in NLP Lab at FI MU (Sedláček, 2001, 2004). Other tools exist for Czech as well (Hajič, 2004) but we prefer Ajka for its properties—it is able to deal with derivational relations automatically.

Ajka is based on the system of the (approx.) 2000 inflectional paradigms, contains about 350 000 Czech stems and is able to generate about 5,7 million Czech word forms. Its coverage/recall for Czech is about 96 % (tested on the corpus All containing 640 mil. Czech word forms and implemented in the NLP Lab at FI MU). It is based on the 'paradigmatic' model of morphology and though it has been primarily devised for Czech its engine can work also with other synthetic languages (such as Slavonic, e.g. Slovak, Serbian, Russian) as well as with analytic ones – there are versions for English, German, French, Dutch, Spanish, Italian.

As we said the morphological module Ajka captures not only the inflectional relations but also the derivation ones (word formation relations). For Czech we know