Polish LFG treebank on a shoestring

Katarzyna Krasnowska[1] Witold Kieraś[1,2]
k.krasnowska@phd.ipipan.waw.pl w.kieras@uw.edu.pl

[1] Institute of Computer Science, Polish Academy of Sciences
[2] Institute of Polish Language, University of Warsaw

Abstract

In the paper we present a method of partial disambiguation of an LFG parsebank produced by the Polish LFG grammar POLFIE. The method is based on the grammatical information retrieved from Składnica treebank consisting of the same set of sentences. As a result we obtain a parsebank consisting of significantly smaller forests of LFG structures that can be fully disambiguated by a human annotator with much less time and effort then in the case of entirely manual disambiguation.

1 Introduction

In this paper, we report on preliminary results concerning a method of semi-automatic creation of an LFG treebank of Polish based on an already existing resource. The aim of our work is to prune LFG forests obtained by automatic parsing with no means of stochastic disambiguation. The idea is to restrict them to trees consistent with already existent constituency annotation based on another grammar formalism for the parsed sentences. In this way, we perform an automatic, partial disambiguation which can be later completed by a human annotator. After the automatic stage of pruning, the annotators will be presented with much smaller parse forests, which will allow for a significant decrease of amount of time and human effort needed to obtain an LFG treebank for Polish.

Our approach is therefore a convenient alternative to more expensive (in terms of time and human effort) ways of creating a treebank, involving fully manual syntactic annotation (e.g., the Prague Dependency Treebank, Hajič et al. 2000) or disambiguation of the entire output of a parser (as in the Polish constituency treebank, Woliński et al. 2011) and therefore requiring more work from the annotators.

The strategy we use, i.e., making extensive use of an existing resource in order to obtain a new one, is in line with the "parasitic" approach adopted by the authors of POLFIE – the Polish LFG grammar (Patejuk and Przepiórkowski, 2012) – who used a Polish DCG[1] grammar as a base for their resource. Such an approach seems

[1] Definite Clause Grammar