How to find the right path?  
(On the morphological disambiguation of sentence in Serbian)  
Cvetana Krstev & Dusko Vitas  
University of Belgrade  
{krstev|vitas}@matf.bg.ac.yu

1 Introduction

On the basis of the method of lexical recognition (Silberztein, 1993) we have developed (Vitas, 2007a) the morphological electronic dictionary of contemporary Serbian (Croatian) in the so called LADL-format (Courtois, 1990). The exploitation of such dictionaries for text and corpus analysis is supported by the programming system Unitex.

The application of the lexical recognition method to the text analysis, unlike the methods based on the machine learning, describes very precisely and with high reliability possible morphological interpretations at the sentence level of the input text. For an arbitrary text less then 1% of all word occurrences remain unrecognized. The obtained results are, however, highly ambiguous in the sense that the word from text (the sequence of alphabetic characters between two separators) can represent several lemmas and/or sets of grammatical categories.

2 An Example

As an example of the mentioned ambiguity we can look at the segment:

(1) Izgledalo mi je da joj više ni do čega nije stalo


“It seemed to me that she cared for nothing anymore”

The graph in Figure 1, which is produced as a result of the analysis by the system Unitex, shows the possible interpretations of this sentence. The number of possible interpretations of the sentence corresponds to the number of paths through the graph multiplied by the number of sets of grammatical categories attached to each node (for instance, the form je can be the form if the verb jesam or the pronoun ona, while the node je,ona has two sets of grammatical categories attached to it – enclitic form in the genitive and the accusative case). It means that the form je can have three possible interpretations. However, except in some exceptional cases, only one path that leads from the starting to the ending node of the graph represents the correct interpretation of the results of the lexical recognition.

---

1 http://infolingu.univ-mlv.fr/english/
3 The Method

The topic of our paper is the investigation of various methods for the automatic detection and elimination of those paths in the graph that are not possible (the 'false ambiguity'). The set of all these procedure we will call the disambiguation procedure. We will discuss the following disambiguation methods:

a. The enhancement of the area in the e-dictionary devoted to the syntactic and semantic markers, for instance, addition of the grammatical features such as natural gender and number for nouns, or semantic features, such as professions.

b. The enhancement of the dictionary with the compound words that would include their inflection (if applicable). We will show that the dictionaries of compound adverbs and prepositions are particularly useful (Krstev, 2006).

c. The stratification of the dictionary. For instance in the example in Figure 1 one potential interpretation of the form \(mi\) is the verb \(miti\) 'to wash' in aorist which is rarely used in contemporary Serbian. With the stratification of the dictionary the verb interpretation of \(mi\) is used if and only if the pronoun interpretation leads to the contradiction.

d. The regular derivation, whose first role is to process the unknown words, such as

\[(2) \quad \text{tridesetrogodišnjakinja}
\]

'thirty three year old woman'

in such a way as to determine its precise grammatical status (Krstev, 2007).

e. Local grammars that enable, for instance, the recognition and tagging of some classes of named entities (Vitas, 2007b).

f. ELAG-grammars (Laporte, 2001) that enable the formalization of the complex conditions for the elimination of the false paths in the sentence graph. These grammars are illustrated by the examples that formulate some agreement conditions for number and gender in Serbian on the basis of the problems explored in (Popović, 2000). Finally, we will discuss how the application of the designed procedure benefits the subsequent syntax analysis, and we will present the limits of the designed procedure.

4 Conclusion

The described procedures enable the recognition of certain number of “false ambiguities” which leads to the removal of a number of paths from the graph. We are trying to answer the question to what extent it is possible to process the results of the lexical analyzer with the presented methods (that are on the level of the local grammars) in order to make them suitable as the input to a hypothetic syntactic parser.
References

Courtois, B. et al. (eds.) (1990) Dictionnaires électroniques du français, Langue française 87, Paris: Larousse


Figure 1 The sentence graph for Izgledalo mi je da joj više ni do čega nije stalo