Univerzita Karlova v Praze

Filozofická fakulta

Ústav teoretické a komputační lingvistiky

Abstrakt disertační práce

Pavel Kopřiva

Automatická identifikace strukturních korespondencí v paralelním korpusu

Automatic identification of structural correspondences in a parallel corpus

Matematická lingvistika — korpusová lingvistika

Vedoucí práce: doc. RNDr. Vladimír Petkevič, CSc.

2013

Abstract (in English)

The aim of this thesis is to design, implement and evaluate an algorithm which will automatically identify corresponding parts of sentences in bilingual parallel texts. These parts are called chunks and their identification is based on recognition of their syntactic structures.

Our proposed algorithm for finding of chunks is based on a set of rules. The rules consist of two main parts: configuration and executive. The configuration part specifies the conditions that must be met so that the rule could be applied. These conditions are the required properties of the chunk sequences in both languages and the initial chunk alignment. The executive part of the rule can then create new chunks, parent–child relationships between them, and align the chunks between the languages. The rules are specific to a given language pair; in our case,

we focused on the Czech and English languages. The algorithm assumes that the input texts are aligned at the word level.

The achieved values of the accuracy measures are not high. Our algorithm was quite successful in identifying chunks and syntactic structures in short sentences. Longer sentences and some grammatical structures are difficult for chunk identification. The results could be improved significantly by increasing the number of rules or by using additional vocabularies. The chunk alignment will contribute to a better understanding of structural similarities and differences between languages and to the improvement of automatic translation.