Lexical similarities among East Slavonic languages as a base for Optimization of Cross-Language Information Retrieval (CLIR) Systems

Natalia Mamerow

Institute of Language and Information, Department of Computational Semantics, Heinrich Heine University Düsseldorf, Germany

Introduction

Research of lexical similarity among genetically related languages provides good starting-points for optimizing CLIR systems. Therefore the east slavonic languages Russian, Belarussian and Ukrainian were choosen for this work. The aim was to find out how high is the rate of cognate words in the three languages.

Methods

The empiric data basis for this research consists of three corpora (each for one east slavonic language) filled with partly parallel texts. The empiric part of this work was limited to the domain „politics“ and the part of speech “nouns”. Data analysis was accomplished with a semi-automatic process. Only nouns were covered by this approach because of their central role as the most important part of speech for information retrieval processing (Losee, 2006, 1258). In the first step all corpora were tokenzied[1]. The extraction of nouns was based on an automatic recognition of flections that are characteristic for east slavonic nouns. In the next step each noun was reduced to its root and the dublicate roots were sort out. In the finalizing step „root-matching“ all roots were compared within all language pairs and among the whole family of East Slavonic languages. The evaluation of similarity occured with a focus on the theoretical possibilities of optimizing natural language processing respectivly CLIR systems using a finite state transducer approach.

Findings

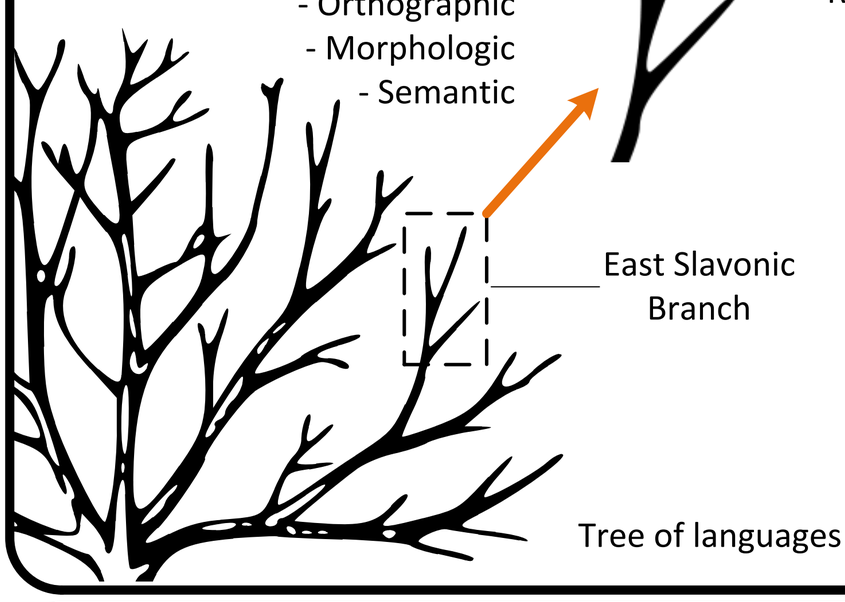
The result of this computer-aided comparison of similarity provides an informative basis about the question how far a better natural language processing for cognates or a CLIR system can be optimized. The empiric analysis showed a maximum potential of 79% for translation on the basis of lexical similarities between Russian, Belarussian and Ukrainian. The results can be integrated in further approaches, e.g. an implementation based on the Finite State Technology for optimizing existing or future information systems. The semi-automatic process for comparing genetic related languages can be adapted for other language pairs, domains and word classes.

1 Adjustment of the tokenizer written by Katina Bontcheva

Focus: East Slavonic CLIR

Cognate comparison

- Orthographic
- Morphologic
- Semantic



Semi-automatic Processing

East Slavonic Corpora

1. Tokenizing

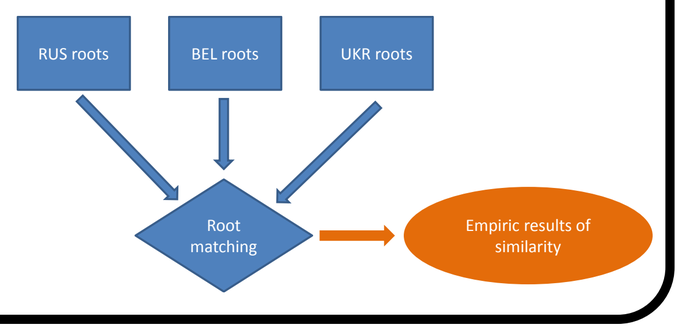
[...] <rv>Конституция Российской Федерации Принята всенародным голосованием 12 декабря 1993 года Мы, многонациональный [...]

Filter out:
Latin characters
Numbers
Special Signs
Puctuation marks
Space characters

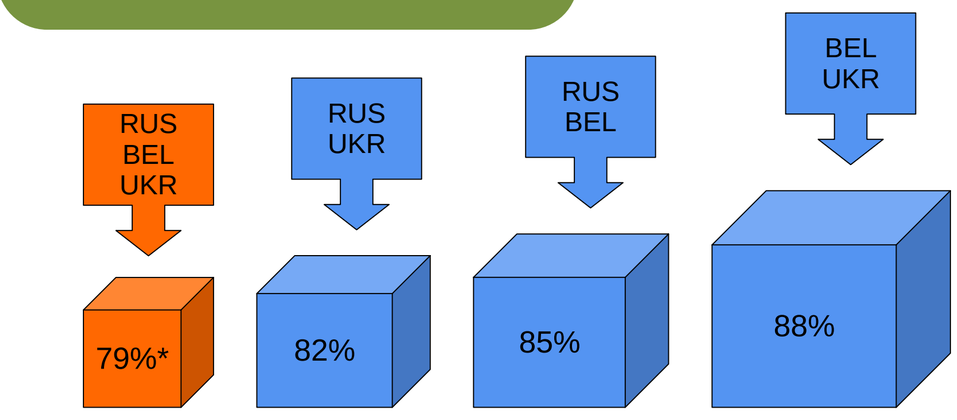
[...] Конституция Российской Федерации Принята всенародным голосованием декабря [...]

2. Extraction & Stemming

3. Root-Matching



Results Overview



*The result for all three languages have to be separated from the individual comparison of each language pair. The potential for CLIR over all three languages is 79%. The potential for bilingual CLIR ranges from 82% - 88% depending on the language pair in focus.

Losee, R. M. (2006). Is 1 noun worth 2 adjectives? Measuring the relative feature utility. *Information Processing and Management*, *42*(5), 1248-1259.