



**Slovenská akadémia vied**  
Jazykovedný ústav Ľudovíta Štúra

# Natural Language Processing, Multilinguality

Sixth International Conference  
Modra, Slovakia, 20–21 October 2011  
Proceedings

Editors  
Daniela Majchráková  
Radovan Garabík

Tribun EU  
2011

**The articles have been reviewed by members of the Program Committee.**

The articles can be used under the  
Creative Commons Attribution-ShareAlike 3.0 Unported License



Slovak National Corpus  
L. Štúr Institute of Linguistics  
Slovak Academy of Sciences  
Bratislava, Slovakia 2011  
<http://korpus.juls.savba.sk/~slovko/>

© by respective authors, 2011  
Editors © Daniela Majchráková and Radovan Garabík, 2011  
Typography © Marek Kukan, 2011  
Cover © Vladimír Benko, 2011  
This edition © Tribun EU, 2011

## Table of Contents

Foreword <i>Mária Šimková</i> .....	7
Úvod <i>Mária Šimková</i> .....	8
Semi-automatic Approach to ASR Errors Categorization in Multi-speaker Corpora <i>Štefan Beňuš, Miloš Cerňak, Milan Rusko, Marián Trnka, Sachia Darjaa and Róbert Sabo</i> .....	9
Towards a multilingual database of verb-related terminology <i>Katarína Chovancová and Jana Klincková</i> .....	18
Bulgarian-Polish-Lithuanian Corpus – Recent Progress and Application <i>Ludmila Dimitrova, Violetta Koseska-Toszewa, Danuta Roszko and Roman Roszko</i> .....	30
Bulgarian–Slovak Parallel Corpus <i>Ludmila Dimitrova and Radovan Garabík</i> .....	44
The Slovak Dictionary of Collocations <i>Peter Ďurčo</i> .....	51
Generating Sets of Synonyms between Languages <i>Ondrej Dzurjov, Ján Genči and Radovan Garabík</i> .....	56
Czech-Slovak Parallel Corpora for MT between Closely Related Languages <i>Petra Galuščáková and Ondřej Bojar</i> .....	65
Slovene Lexical Database <i>Polona Gantar and Simon Krek</i> .....	72
Building Annotated Corpora without Experts <i>Marek Grác</i> .....	81
The Lexicographic Representation of Czech Diatheses: Rule Based Approach <i>Václava Kettnerová and Markéta Lopatková</i> .....	89
Translating prepositions from Czech into Russian: challenges for the Machine Translation <i>Natalia Klyueva and Naděžda Runštuková</i> .....	101
Extracting Verbs with PP/NP Variation from the Large 3-gram Corpus <i>Mikhail Kopotev, Natalia Kochetkova and Eduard Klyshinsky</i> .....	109

On the Question of Homonymy and Polysemy in the Lexicographical Practice of the Russian Language Semantic System in its Development Modeling <i>Irina Nekipelova</i> .....	115
Historical and Etymological Electronic Dictionary System Modelling <i>Irina Nekipelova and Elvira Zarifullina</i> .....	121
A Web-based Morphological Tagger for Bulgarian <i>Aleksandar Savkov, Laska Laskova, Petya Osenova, Kiril Simov and Stanislava Kancheva</i> .....	126
Automatic Extraction of Multiword Expressions using Linguistic Constraints for Slovak LVCSR <i>Ján Staš, Daniel Hládek, Marián Trnka and Jozef Juhár</i> .....	138
Common Formal Framework for Multilingual Representation of Inflectional Morphology for Two Related Slavonic Languages <i>Velislava Stoykova</i> .....	146
Recent Developments in ParaSol: Breadth for Depth and XSLT based web concordancing with CWB <i>Ruprecht von Waldenfels</i> .....	156
The Instrumental Environment for the Automatic Syntactical Analysis of Ukrainian <i>Iryna Zamaruieva and Olga Shypnivska</i> .....	163

## Foreword

After two years since the last International Conference under the umbrella name Slovko, we now meet at the 6<sup>th</sup> edition of the event which is primarily focused on Natural Language Processing (NLP) and multilingualism. According to the proposals submitted, authors of the contributions focused mainly on the issues of parallel corpora and automatic translation, corpus based grammar research and occasionally other discussed issues. There is an absence of lexicographers and terminologists, much like the e-learning subject, which seems to be less prevalent in the academic and university scene whose representatives regularly take part in the Slovko conference.

While organizing previous editions, the main theme and event venue in particular were at the centre of attention. The focus of the following edition is already taking its shape: semantic networks and multilingual dictionaries. However, under different circumstances, other questions have come up which have led to certain organizational changes. This will be reflected in the form of the published contributions in a consistent way as to formally address the adjusted requirements for publications evaluating.

The 6<sup>th</sup> conference edition will take place in a small city called Modra famous for its fine wine and blue-and-white porcelain as well as for the renowned linguist Ľudovít Štúr, author of the Slovak orthography rules and whose name our institute also bears. One of his visions was permanent Slavic collaboration and six of the Slavic nations will be present at the conference (Bulgaria, Czech Republic, Slovakia, Slovenia, Russia, and Ukraine). Naturally, our events are open to all languages (Štúr himself could speak several Slavic languages plus Latin, German, Hungarian, French, Greek, and was studying Hebrew and English) and at the conference, more languages used to be present. Each edition is different and each time it brings new preferred issues, presenters and discussants. We are able to meet scientists from other countries in various forums while dealing with projects on NLP, Language Resources, and Corpus Linguistics.

To all the participants of the 6<sup>th</sup> Conference edition, we hope you have a pleasant stay in Modra and enjoy fruitful discussions and mutual inspirations that will continue beyond this meeting and become a part of prospective cooperation.

*Mária Šimková*  
*Translated by Adriána Žáková*

## Úvod

Po dvoch rokoch od posledného stretnutia na medzinárodnej konferencii so zastrešujúcim názvom Slovko sa stretávame na 6. ročníku tohto podujatia, ktoré je primárne zamerané na počítačové spracovanie prirodzeného jazyka a mnohojazyčnosť. Záujem autorov príspevkov sa podľa návrhov predložených v propozíciách sústredil najmä na paralelné korpusy a otázky automatizovaného prekladu, významne sú zastúpené gramaticky orientované korpusové výskumy, ojedinele i ďalšie z ponúknutých tém. Citeľná je neprítomnosť lexikografov a terminológov, bez zastúpenia ostala aj téma e-learningu, zrejme všeobecne málo rozšírená na prevažne akademických a univerzitných pracoviskách, ktorých zástupcovia sa pravidelne zúčastňujú konferencie Slovko.

Pri organizovaní doterajších ročníkov sme si kládli predovšetkým otázku hlavnej témy a miesta konania podujatia. Zameranie nasledujúceho ročníka sa nám už črtá: sémantické siete a multilingválne slovníky. Súčasne sa však tentoraz, pod vplyvom rôznych okolností, vynorili ďalšie otázky, ktoré nás nabádajú k istým organizačným zmenám. Najviac sa to prejaví na spôsobe publikovania príspevkov tak, aby to zodpovedalo formálne nastaveným kritériám na hodnotenie publikačných výstupov.

6. ročník konferencie Slovko sme umiestnili do malého mestečka Modra, ktoré je okrem dobrého vína a kvalitnej keramiky známe aj pôsobením kodifikátora spisovnej slovenčiny Eudovíta Štúra, ktorého meno nesie naše kmeňové pracovisko. Jednou z jeho vízií bola spolupráca slovanských národov, z ktorých šesť je zastúpených aj na tomto stretnutí (Bulharsko, Česká republika, Slovensko, Slovinsko, Rusko, Ukrajina). Naše podujatia sú, samozrejme, otvorené voči všetkým jazykom (aj sám Štúr ovládal okrem viacerých slovanských jazykov ešte latinčinu, nemčinu, maďarčinu, francúzštinu, gréčtinu, učil sa hebrejčinu a angličtinu) a na Slovku ich priebežne bývalo zastúpených viac, ale každý ročník je iný, prináša vždy aktuálne nové preferencie tém a nové zloženie prezentujúcich a diskutujúcich. S predstaviteľmi ďalších krajín sa zas stretávame na iných fórach a pri riešení projektov v oblasti počítačového spracovania jazyka, jazykových zdrojov a korpusovej lingvistiky.

Účastníkom 6. ročníka konferencie Slovko želáme príjemný pobyt v Modre, plodné rokovania a vzájomné inšpirácie, ktoré prekročia rámec tohto stretnutia a stanú sa súčasťou perspektívnej spolupráce.

*Mária Šimková*

# Semi-automatic Approach to ASR Errors Categorization in Multi-speaker Corpora

Štefan Beňuš<sup>1,2</sup>, Miloš Cernák<sup>1</sup>, Milan Rusko<sup>1</sup>, Marián Trnka<sup>1</sup>, Sachia Darjaa<sup>1</sup> and Róbert Sabo<sup>1</sup>

<sup>1</sup> Institute of Informatics, Slovak Academy of Sciences, Bratislava

<sup>2</sup> Faculty of Arts, Constantine the Philosopher University, Nitra

**Abstract.** Error diagnosis is an integral part of improving the quality and robustness of any ASR system, especially for languages with limited resources. This paper explores a semi-automatic approach to error categorization usable for databases that have a set of identical sentences produced by a sufficiently large number of speakers. We use a matrix created from an ordered list of speakers and an ordered list of sentences based on the recognizer performance. An algorithm that searches through the errors using such a matrix is proposed and the utilization of information obtained from the output is discussed.

## 1 Introduction

Error diagnosis is an integral part of improving the quality and robustness of any ASR system. It is even more important for languages with small population and limited resources since increasing the size of the training data or adapting to specific domains is more challenging than for languages with less limited resources. Moreover, good understanding of the nature of ASR deficiencies provides useful information for the type of data collection most likely to improve the performance of the system and thus facilitates the efficient allocation of the resources. In this paper we describe a methodological approach to a systematic analysis of ASR errors that is usable for databases that have a set of identical sentences produced by a sufficiently large number of speakers. Our approach is based on a semi-automatic analysis of a matrix created from an ordered list of speakers and an ordered list of sentences based on the recognizer performance.

When speech recognition systems leave laboratory conditions and are used in real life, they are faced with error handling everywhere, because recognition errors are ultimately unavoidable. Error handling is typically more in the interest of application architects than of ASR researchers; for example, in applications such as dialog systems, errors are usually handled at higher levels of the application [1].

In ASR application without a human-machine communication module, which is a situation similar to our efforts at building a reliable ASR system for recognizing and transcribing dictated or spoken Slovak, Nanjo et al. [2] experimented with diagnosing recognition errors in three tasks: a) 5K-word and 20K-word dictation of Japanese newspapers, 2) lecture speech transcription system, and c) dialogue speech recognition system. Based on the evaluation of partial acoustic and language model probab-

ities for recognized and correct word sequences, their fully automatic categorization procedure classified errors into four main categories (acoustic model (AM) errors, language model (LM) errors, decoder errors attributable to problems in searching for the best hypothesis combining the AM and LM scores, and errors linked to both the acoustic and language models). This approach proved useful primarily for improving the decoder performance. Errors attributable to the lexicon deficiencies were not considered.

Our goal in this paper is to improve our understanding of the error patterns and generate knowledge that would be applicable towards the improvement of the accuracy of our state-of-the-art ASR system for Slovak. We are interested in identifying these potentially problematic areas: identification of word/pronunciation variants that are not covered by the lexicon, language model (LM) problems especially due to back-off smoothing, acoustic model (AM) problems related to intra- and inter-word triphone contexts, and problems caused in the segmentation of compounds and morphologically-complex words.

Given the fact that one of the relatively common features of speech databases is a set of identical sentences produced by multiple speakers, our semi-automated approach may complement more sophisticated automatic methods of error diagnostics or produce a first crude pass through the error space that facilitates further analyses.

## 2 Method and data

### 2.1 Testing corpus

Our testing corpus consists of 380 sentences read by 18 native speakers of Slovak recorded as a part of building a larger corpus for automatic applications to second language learning.

These sentences were specially selected from existing texts to create a phonetically rich and balanced corpus in the following way. An iterative algorithm searched in the available corpus for sentences with specified word count of one to seven words. The greatest emphasis was put on collecting sentences with the greatest number of non-identical phonemes. Additionally, the algorithm preferred sentences with triphones poorly represented in the already selected pool of sentences. Once a sentence was put into this pool, the triphone statistics were re-calculated and the search continued for the next sentence until the target number of sentences was selected. Table 1 below displays the word and triphone coverage in the 380 selected sentences.

Words		Triphones		intra-word triphones		inter-word triphones	
all	unique	all	unique	all	unique	all	unique
1 970	1 739	17 302	5 174	12 898	4 043	4 404	1 849

**Table 1.** Word and triphone coverage of the corpus

## 2.2 Automatic speech recognition system for Slovak

Experiments have been performed using read, spontaneous, and broadcast speech databases of Slovak language. The first database (D1) contained 250 hours of gender balanced read speech, recorded from 250 speakers with Sennheiser ME3 Headset Microphone with In-Line Preampfier Sennheiser MZA 900 P. The second database (D2) contained 100 hours of 90% male spontaneous speech, recorded from 120 speakers at council hall with goose neck microphones. The third database (D3) contained 200 hours of broad cast data selected primarily from two Slovak TV stations (STV and TA3), encoded with MPEG-2 and resampled to 16 kHz. Databases were annotated using Transcriber annotation tool [3], twice checked and corrected. Recordings were split on segments if possible not bigger than 10 sec. Special labels, studio environment and telephone speech, were used in broadcast data transcription.

We experiment with three ASR systems trained with progressively more data. The first experiment (EX1) used a subset of the read speech database D1 that contained 130 hours recorded by 140 speakers. The second experiment (EX2) used both read and spontaneous speech databases (D1+D2; 350 hours recorded by 370 speakers), and in the third experiment (EX3) broadcast data has been included. Julius decoder [4] was used as a reference speech recognition engine, and the HTK toolkit was used for word-internal acoustic models training. We trained AMs using the triphone mapping as described in [5], with 32 Gaussian densities per each HMM state.

The text corpora contained a total of about 92 million sentences with 1.25 billion Slovak words. Trigram language models (LMs) were created with a vocabulary size of 350k unique words (400k pronunciation variants) which passed the spell-check lexicon and subsequently were also checked manually. As a smoothing technique the modified Kneser-Ney algorithm was used [6].

## 2.3 Output processing

The output of each of the three ASR systems is then processed as follows. We determine an ordered list of speakers and an ordered list of sentences based on the recognizer performance. For the sorting measure we use average word error rate (WER) in case of the speakers and the sum of insertion, deletion, and substitution errors divided by the number of words in case of the sentences. Informally, we will refer to speakers with low WER as ‘good’ speakers and with high WER as ‘bad’ speakers, and sentences with low error metric as ‘easy’ sentences and those with high error metric as ‘difficult’ sentences. It should be noted, however, that these descriptor words do not characterize speakers or sentences themselves but rather the degree of difficulty they pose for our ASR system.

In this way, and despite the fact that both lists present continua rather than binary distributions, the four corners of a two-dimensional matrix created from our two ordered lists can be characterized as 1) easy sentences by good speakers, 2) difficult sentences produced by bad speakers, 3) easy sentences and bad speakers, and 4) difficult sentences and good speakers. This arrangement provides us with first crude

information about the performance of the system. For example, sentences correctly recognized for all 18 speakers point to the areas of robustness of the acoustic model giving 100% correct recognition despite inter-speaker variation in voice quality, speech rate, prosodic patterns, and speaking style.

Next, we categorize the errors into three broad types: 1) those involving only a substitution such as *pristrihnú* recognized as *pristrihnúť* or *pohltil* recognized as *pohltila*, 2) those in which a deletion or an insertion occurs together with, that is, adjacent to, a substitution such as *posudzovateľmi* recognized as *posudzovateľ mi* or *pokývávali* recognized as *pokým cavalli* or *pokým káva ich*, and 3) miscellaneous errors in which the recognized word did not correspond to a reference word but could not be assigned to the first two categories.

The rationale for this categorization is that in a language with rich inflectional morphology like Slovak, the first type of errors typically signals a word with a correctly recognized stem but erroneous inflectional affix. The second type of errors commonly involves either a split of a problematic reference word into two substrings, which corresponds to the presence of a substitution error adjacent to an insertion error, or joining of two reference words into a single string, which produces a substitution error adjacent to a deletion error. These errors then seem to pertain to the balance in the inter- and intra-word triphones in the AM. This is because a split of a reference word into two recognized words incorrectly prefers an inter-word triphone over an intra-word one. Alternatively, joining of two reference words into a single recognized word signals an incorrect preference for intra-word triphone over an inter-word one.

Using this information, the next step of our error diagnosis routine includes populating the cells of the matrix with the individual errors and error types. Processing the matrix sentence-wise, the most obviously useful information we obtain is the identification of those errors in which a reference word was incorrectly recognized for all 18 speakers. These errors are most likely attributable to the lexicon deficiency or the LM model. Additionally, each error can be further weighted based on how good a speaker producing the error was and by how many speakers produced this error:

$$w = \frac{a}{N} \sum_i b_i \quad (1)$$

where  $a$  is the number of speakers producing the error normalized by  $N$ , the number of all speakers, and  $b$  is the rank of the speaker producing the error; with  $i$  looping over all speakers that produced the error.

In this way, if errors by good speakers are weighted more than those from bad speakers, and more frequent errors are weighted more than less frequent one, a resulting index can be construed as corresponding to the likelihood that a particular error can be attributed to the lexicon or LM. Hence, if an error is relatively frequent, and it is commonly produced by good speakers, its likelihood of belonging to the AM component decreases.

As a final step in the processing of the errors we tried to estimate the location of the error. More specifically, we coded if the error occurred on the left edge of the reference word, on the right edge, in the middle, or if this information cannot be determined. The algorithm takes the pairs of reference and hypothesized words (or sequences), aligns them based on graphemes, calculates a measure of distance between the two strings, and if the two sequences are sufficiently similar, typically 70% and more, the algorithm attempts to determine the location of the error within the sequence. This information is especially useful for substitution errors since the left or the right edge location points to a likely inter-word problem. One way of approaching this information is to extract the relevant context of the reference word, hence, the word preceding and/or following the reference word, and then investigate the effect of this context as a potential trigger. For example, in *čítam pekne* recognized as *číta pekne* or *čítam iba* recognized as *čítaj iba* we determine that the problem occurs on the right edge of the first reference word, extract the word following the first reference word, and we can look for patterns in the observed problematic inter-word triphones such as a-m#p and a-m#i.

### 3 Results and discussion

#### 3.1 Comparing three ASR systems

The table below shows the performance of our three ASR systems described in section 2.2. It is interesting that while the addition of read speech together with the speech from parliamentary speeches improved the accuracy of recognition, as can be seen by comparing EX1 and EX2, further addition of broadcast data caused the deterioration of the performance by 2%. This is somewhat unexpected since the testing corpus involves reading of sentences, which might seem a similar domain to reading news on television and radio (but see below).

ASR	WER
EX1 (D1)	22.3
EX2 (D2)	21.33
EX3 (D3)	23.63

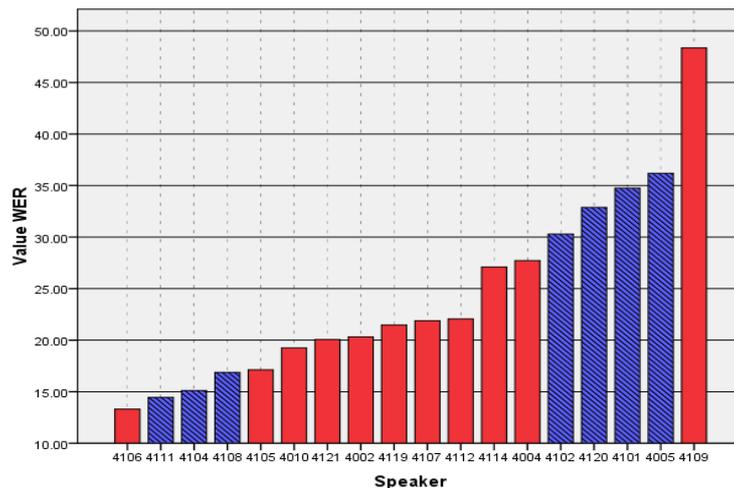
**Table 2.** Performance of three tested ASR systems on the test corpus

Two potential reasons for this deterioration are noisy data, and different prosody patterns. First, as explained in section 2.2, we employed automatic measures that identify the chunks of broadcast speech that occur in the studio and exclude telephone speech or speech in noisy environment. However, it is possible that despite these measures, the broadcast database still includes section of noisy data. Second, it is known that news broadcasters employ rather different prosodic patterns for read sentences as compared to natural spontaneous narratives in an effort to increase the dynamic tension of each sentence and maintain the attention of viewers between

sentences. These strategies mostly relate to the placement of the pitch accents and phrasing of sentences into smaller prosodic domains; e.g. [7]. Another issue is that broadcast speech is quite fast given the fixed time constraints while the subjects in our corpus had no time pressure for reading the sentences and were thus presumably much slower than broadcast anchors.

### 3.2 Descriptive analysis of observed error patterns

For the error analysis we selected the best-performing ASR system, that is, the one we refer to as EX2 in Table 2 above and section 2.2. The comparison of the errors based on the three recognizer systems is beyond the scope of the current paper and we leave it for future research. The first step of our algorithm created a two-dimensional matrix of the ASR output. Figure 1 shows the distribution of speaker's WER ranging from 13.3 for the best one to 48.4 for the worst one. The distribution is fairly continuous but two minor discontinuities can be observed between the 11<sup>th</sup> and the 12<sup>th</sup> speaker (4112 and 4224 respectively) and before the worst speaker (4109). We can also observe somewhat surprising clustering of speaker gender.



**Fig.1.** Ordered word error rate (WER) of individual speakers from EX2 system split by gender (red filled bars correspond to females).

This first processing (the matrix of ordered lists that was created following the steps described in section 2.3) showed that only 7 sentences (1.8%) were correctly recognized for all 18 speakers. These were not particularly short; six contained 4 words and one 5 words, which corresponds to the average length of sentences in the testing corpus. This result suggests that the domain of the highest robustness of our recognizer is relatively narrow and that knowledge gained from error diagnosis and analysis might bring significant improvements in the robustness of our ASR system.

The subsequent processing of errors and error types based on the matrix structure and described in 2.3 revealed several notable and systematic patterns that we discuss in the remainder of this section.

First, the mismatch between the spelling conventions in the training and testing corpora accounted for a significant number of non-AM errors, especially those linked to numerals. While the reference texts had complex numerals written as one word, e.g. *dvadsaťtri*, the training corpus gave greater weight to these numerals written as separate words, e.g. *dvadsať tri*. A similar pattern was observed for several compound words such as *hrôzy-plný*, *troj-kilometrový*, or *zoči-voči*. In terms of assessing the accuracy of the recognizer in a particular application, these types of errors, in which the compound reference word is completely identical to the string created by joining the separately recognized parts of the compound, might be considered as minor or negligible errors.

Second, our ASR system used in the experiments does not contain the module differentiating intra-word triphones from the inter-word ones. However, a large percentage of errors attributable to AM occurred at the edge of the words, especially the right edge. The most common were those related to phonetically similar and extremely frequent suffixes. For example, *-n-e* [ɲe] vs. *-n-ie* [ɲie] differentiates adverbs and adjectives in neuter gender, the addition of *-t'* [c] differentiates the verbal infinitives from 3<sup>rd</sup> person present tense forms, or the addition of *-m* contrasts the first person singular from the 3<sup>rd</sup> person. Both consonantal suffixes, i.e. *-m* and *-t'*, when followed by words starting with consonants of a similar manner or place of articulation, create the environment for recognition problems. This is a complex issue since the right edge of the target word is the site of multiple derivational morphemes that have inherent imbalance in the frequency of their distributions. However, the right edge of each word is the site of pervasive regressive coarticulation that tends to affect the word-final sounds more than the word-initial ones.

Consider another example of highly frequent errors attributable to AM as shown in our diagnostic procedure; there are four past tense verbal suffixes in Slovak: *-l*, *-la*, *-lo*, *-li*.<sup>1</sup> The high frequency of errors associated with these endings is due to the acoustic similarity of /a/ and /o/, the similarity of /l/ to both of the vowels, the coarticulation of word-final vowels with word-initial sounds, or shortening of word-final vowels especially if the following word start with a stop consonant. Similar problems occur also with other frequent suffixes. We plan to test if these types of word edge problems might be alleviated through 1) introduction of a subset of the most frequent inter-word triphones into our training routines, and 2) the introduction of a module that would produce a shallow parts of speech tagger together with rudimentary stemming of each word. The module would then check for the grammatical agreement of some adjacent categories with well-defined agreement rules (e.g. adjective and noun, or preposition and noun).

---

<sup>1</sup> Actually, these suffixes are further decomposable into a past tense morpheme *-l* and then zero ending for masculine, *-a* for feminine, *-o* for neuter, and *-i* for all plural forms.

Third, a rather surprising problem was identified at the left edge of the words. In addition to expected problems with prepositions, that are frequently identical to prefixes, we also identified an unexpected problem with a common prefix *ne-* [ɲe] that signals negative polarity of verbs. An unusually common error involved a negative-polarity verbal form being split into a phonetically identical, but highly infrequent particle *ne* as a separate word followed by the verb without the negative-polarity prefix. For example, *nepríjemnejšiu* was recognized as a grammatically illicit bigram *ne príjemnejšiu*. Furthermore, other more frequent and phonetically similar words such as *je* [je] ‘is’, *nej* [ɲeɪ] ‘her’ or *mne* [mɲe] ‘me’ could replace the negative polarity prefix in verbs.

Fourth, the last systematic error included the recognition of spelled letters as part of reference words. For example, *riadiateľove laboratória* was recognized as *riadiatelia V laboratória*, or *v krízových situáciách* was recognized as *kríza ich C dva C tri L*. For our dictation ASR recognition we are experimenting with a user mode for spelling proper names, which then allows us to weigh spelled letters in the recognized text lower, hopefully significantly decreasing the errors stemming from the inclusion of spelled letters in the recognizer output.

All four of the above issues are systematic in the sense that the identification of the problem is not linked to the testing corpus only, but is extendable to problems encountered in real-world applications of the recognition system. The final issue we mention is pervasive, yet non-systematic. Our approach identified multiple relatively infrequent words absent or wrongly spelled in the lexicon such as *neoveriteľných*, *nevyhľadávaj*, *čalúnnikov*, *rozhadzovačnosťou*, *ikonografický*, and many others. Obviously, these errors are closely linked to the rare words contained exclusively in the training corpus. Therefore, their treatment would significantly improve the accuracy of our experiments but would likely result in negligible gains when the ASR system is deployed in real-world applications.

Finally, additional problematic issues that would require further processing of the matrix but are also approachable in the following way include the detection of the initial trigger of the recognition errors that have a great impact on the success of recognizing the subsequent words in the sentence. A potential approach could include weighting errors occurring in the reference words from left to right in the sentence ranking the ones toward the beginning of the sentence more than those toward the end of it. Additionally, in a sequence of words  $w_1 w_2 w_3$  a substitution error in  $w_1$  (ET1), or a substitution plus insertion error (ET2) would receive greater weight if the subsequent words  $w_2$  and/or  $w_3$  were incorrectly recognized. Knowing the distribution of these weights together with the error types and combining that with the location of the error in the error matrix provides further useful avenues for analysis and treatment, for example in the area of triphone training. We also plan to experiment in the future with comparing the error measures of multiple ASR experiments (such as the ones reported in Table 2) with the hope that the error metrics might provide complementary information to the standard metrics such as WER.

## 4 Conclusion

In this paper we explored a semi-automatic approach to error categorization in the output of ASR system for Slovak. Our approach is usable for databases that have a set of identical sentences produced by a sufficiently large number of speakers. We suggested utilizing a matrix created from an ordered list of speakers and an ordered list of sentences based on the recognizer performance. We discussed a step-wise process that searches through the errors using such a matrix and provides useful information at each step of the process. This way of analyzing errors provides us with multiple sources of information such as the probability that the error is attributable to the acoustic model, and the identification of whether errors can be characterized as problematic due to inter- or intra-word triphone training.

## References

- [1] Authors Editorial (2005). Error handling in spoken dialogue system, *Speech Communication* 45, 207-20.
- [2] Nanjo, H., Lee, A., Kawahara, T. (2000). Automatic Diagnosis of Recognition Errors in Large Vocabulary Continuous Speech Recognition System. In *Proceedings of International Conference on Spoken Language Processing*.
- [3] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. (2000). Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33 (1–2), 23–60.
- [4] A. Lee, T. Kawahara, and K. Shikano. (2001). Julius – an Open Source Real-Time Large Vocabulary Recognition Engine. In *Proc. of the European Conference on Speech Communications and Technology (EUROSPEECH)*.
- [5] Darjaa, S., Cerňak, M., Trnka, M., Rusko, M., Sabo, R. (2011). Effective Triphone Mapping for Acoustic Modeling in Speech Recognition. Ms, submitted, Intrespeech 2011.
- [6] Staš, J., Hládek, D., Juhár, J. (2010), Language Model Adaptation for Slovak LVCSR. In *Proceedings of the International Conference on AEI*, pp. 101–106.
- [7] Shriberg, E., Favre, B., Fung, J., Hakkani-Tür, D., Cuendet, S. (2009). Prosodic Similarities of Dialog Act Boundaries Across Speaking Styles. *Linguistic Patterns in Spontaneous Speech*, 213-239.

# Towards a Multilingual Database of Verb-related Terminology

Katarína Chovancová and Jana Klincková

Matej Bel University, Banská Bystrica

**Abstract.** Contrastive studies of linguistic metaterminology responds to an urgent need of the academic and professional community, which often encounters problems while choosing appropriate terms to describe linguistic phenomena. This paper presents a contrastive analysis of verb terminology in Slovak and in Romance languages with the aid of electronic databases.

## 1 Introduction

Terminology is a pillar of communication within the sciences. Creating, searching for and using texts that contain units of knowledge (in the form of concepts and terms) is of great importance for the further development of scientific disciplines and domains of human activity. Development of the knowledge-based society and contacts between cultural and linguistic communities is accompanied by a constant need to communicate specialized information in different languages. This is reflected in an increasing interest in the methods and principles of terminologic work and in the development of terminologic databases as starting points for the consistent repeated use of terminologic data.

Contrastive studies of linguistic metaterminology such as ours are not often found at the top of the list when it comes to supporting terminologic research. Still, it responds to an urgent need of the academic and professional community, which often encounters problems while choosing appropriate terms to describe linguistic phenomena. This paper presents a contrastive analysis of verb terminology in Slovak and in Romance languages with the aid of electronic databases.

## 2 The theoretic and methodologic bases of terminologic research

### 2.1 Classification of concepts and modelling of conceptual systems

A term refers to a concept. A term is determined by its definition and its place within the conceptual system of the given discipline. Whereas a concept conveys extralinguistic information; it gives name to an extralinguistic phenomenon. A term is a part of the language lexicon; it is a linguistic sign. A linguistic term is an element of metalanguage; it establishes an intralinguistic relationship. The conceptual content of a linguistic term depends on theoretic approaches to the study of language (Ibrahim – Křečková, 2009).

Each field of knowledge has its own concepts, which in turn form systems and subsystems. Within these, each concept occupies its own specific position. The classification of concepts and modelling of conceptual systems are the basis of terminologic research. They reveal the relationships between concepts and terms. The classification of concepts reflects the current state of the field and results from analyses of classified objects.

The compilation of multilingual terminologic sets begins with modelling a conceptual system for each represented language, taking into account national variations and various theoretical approaches. The main classificatory criteria should be the same for each language. Modelling conceptual systems is a complex, yet crucial part of terminologic work, as it sets up a basis for the equivalence of concepts and terms. In linguistics, the process is complicated even more by the distinctive features of linguistic terminology.

## 2.2 The possibility of harmonizing conceptual systems

In different language communities, concepts and terms are perceived in various ways, depending on social, economic, cultural and linguistic factors. Differences between parallel conceptual systems can lead to problems in international communication. Increasing communication needs and the development of international contacts in the sciences and technology are thus accompanied by the necessity to harmonize concepts and terms<sup>1</sup>. To harmonize terms means to design the same concept in various languages by terms that reflect the same or similar features and have the same or similar forms. Harmonization should make conceptual systems match. This objective is reached through comparing definitions. Harmonization of terms, concepts and conceptual systems can lead to problems in certain fields of knowledge. Conceptual systems differ when there are disparate theoretic approaches in the given domain or if the field in question has no tradition of harmonization. These criteria are stated in the ISO 860 standard and concern, among other fields, linguistics and linguistic terminology. Given the particularities of linguistic terminology, harmonization of concepts and conceptual systems within this field is not an easy task.

Marc Van Campenhoudt (1996), referring to Reiner Arntz<sup>2</sup>, recommends establishing equivalence by comparing the various conceptual systems with which languages are used to describe themselves. This comparison will expose the differences among conceptual representations of the field in various languages, while achieving interlingual equivalence will result from conceptual analysis of terms in each lan-

<sup>1</sup> The Vienna terminologic school (Eugen Wüster, Helmut Felber) proposed harmonization and unification of conceptual systems on an international level (see Van Campenhoudt, 1996). The idea gave birth to the international terminology standard ISO 860, translated into Slovak by Z. Jurčáková and E. Krištofičová and published with the title *Terminology work: Harmonization of concepts and terms* (ISO 860 : 1998).

<sup>2</sup> ARNTZ, R. 1993. Terminological Equivalence and Translation. In SONNEVELD, H. – LOENING, K. (ed.) *Terminology. Applications in Interdisciplinary Communication*, Amsterdam – Philadelphia, John Benjamins Publishing Company, pp. 5-19.

guage (the semasiologic approach). The equivalent in another language will in turn be obtained by applying an onomasiologic approach. Such a procedure will show the differences among the conceptual structures of terms in various languages and address cases of partial equivalence using translation methods.

Pierre Lerat (2010) analyses the problem of variability and harmonization of terms and concepts while creating multilingual terminologic databases. He also affirms that harmonization is not always easy. The ISO 860 standard recommends formulating a definition of the term using the same characteristics in each language. Still, the formulation of definitions depends at the same time on the rules of each language. In a multilingual terminologic database that shows the differences among various languages' conceptual structures of terms, it is recommended to make a distinction between two groups of information: the description of the term by a definition and by its position in the conceptual system of the subject field.

When establishing equivalence between two different languages, the term is represented in each of them not only by its basic definition, but also by a note which supplements the definition and shows differences the various conceptual systems with which languages are used to describe themselves among the languages' conceptual structures. Terminologic variability in the creation of terminological sets can be solved, according to Lerat (2010), by establishing a referential term, of which competing terms are then considered synonyms.

In the field of multilingual terminology, where harmonization of concepts and terms is difficult to accomplish, equivalence can be established when terms are defined on the basis of the same characteristics. A linguistic concept – apart from those referring to language universals – refers to one language; complete equivalence of concepts and terms in the field of linguistics is rare indeed.

### 2.3 Discordance among conceptual systems of grammar metalanguage

One example of discordance between linguistic terms is the Slovak term *limitné sloveso*, reflecting differences in the perception of a concept related to verb classification.

*Limitné sloveso* is a term used in Slovak linguistics to describe those verbs that 'v spojení s plnovýznamovým slovesom vyjadrujú konečný prípravný stav pred uskutočnením deja. Majú istý vzťah ku gramatickej kategórii času. Spojenia týchto pomocných slovies s neurčitkom dokonavých slovies vyjadrujú, že vyslovený dej sa už-už má uskutočniť, alebo len-len že sa neuskutočnil.'<sup>3</sup> (Morfológia slovenského jazyka, 1966, p. 365; similarly Navrátil, 2009, p. 24-25; Sokolová, 1993, p. 21) *Limitné slovesá* (plural form) include *mať* (to have) and *ísť* (to go), e.g. *malo ho rozdrapit'* (he was about to burst out), *išlo ho poraziť* (he nearly had a heart attack), *nejdem sa rozčuľovať* (I'm not gonna lose my temper), *zdalo sa, že ide plakať* (he appeared close

---

<sup>3</sup> In relation to a full verb, they express the final preparatory state before the realization of the action. They are related to the category of tense. Collocations of auxiliary verbs and infinitives of perfective verbs are used to express that the action is about to happen or almost happened.

to tears). These examples show that to define the term *limitné sloveso*, several semantic characteristics have to be taken into account, especially the fact that these verbs indicate the state immediately before the realisation of the action (the end of the action's preparatory phase), as well as the fact that the action is not realized. When it comes to the semantic identification of verbs, the values of *íst'* are distinguished according to infinitives that complete its meaning (cf. *idem sa učít' do študovne* (I'm going to the study to study) = 1. *idem do študovne, aby som sa učila* (I'm going to the study room in order to study), 2. *práve sa idem učít' do študovne* (I'm going to the study to study at the moment) and according to its expressivity; it is used to express the near future (*ide snežiť'* (it's about to snow), *čo ideš robiť'* (what are you going to do now)) or the limited character of the realization of the action (*mala som sa popučít' od smiechu* (I was about to burst out laughing), *ide ho rozhodít'* (he's about to explode). (Sokolová, 1993, p. 85)

At first sight, French linguistic literature does not offer an appropriate equivalent of *limitné sloveso*. Examples such as *malo ho / išlo ho rozdrapít'* are translated into French as

- (1) il était sur le point d'exploser
- (2) il allait exploser
- (3) il a failli exploser

Upon closer inspection, we see these expressions' variable character. The first two cases reveal themselves to be periphrastic descriptions of verb tense, as in the expression *être sur le point de* + infinitive (ex. 1), which belongs to the French category of *périphrases verbales*. In example 2, we see the expression of a near future in the past, using *aller* + infinitive. According to Riegel et al. (1994, p. 253), both expressions belong to the category *auxiliaires d'aspect*, the group of auxiliaries expressing slight nuances of an action. Other expressions of this kind include *être en passe de*, *commencer à/de*, *se mettre à*, *être en train de*, *aller* + *participe présent*, *finir de*, *cesser de*, *achever de* and *venir de*. French *auxiliaires d'aspect*, together with *auxiliaires modaux* (Riegel speaks namely of *pouvoir* and *devoir*, as well as some uses of *paraître*, *sembler* and *vouloir*) are designated as *semi-auxiliaires*. Verbs classified as *auxiliaires d'aspect* allow one to express a series of meanings related to phases of the action. Expressions such as *aller* + *infinitif*, *être sur le point de* and *être en passe de* are used to express the moment that immediately precedes an action's actual realisation. While this definition corresponds to the Slovak definition of *limitné slovesá*, the French nonetheless lacks a specific term to delimit verbs of this kind within the group of *verbes auxiliaires* or *semi-auxiliaires*<sup>4</sup>.

The third example shows the verb *faillir*, which belongs among *verbes semi-auxiliaires* as well (Grevisse, 1969, p. 599). This verb is used to underline the fact that the action has not yet been realized. We can thus paraphrase: *Il a failli exploser, mais il ne l'a pas fait* (He was about to explode, but he didn't.)

<sup>4</sup> In Slovak, with regard to the close relation of *limitné slovesá* to the category of tense, they behave as grammar functors of temporality (Navrátil, 2009, p. 25), or they are considered as modifiers (Kralčák, 2005, p. 48) or limiting modifiers (Nižníková – Sokolová, 1998, p. 8).

*Limitné slovesá* are an example of the situation when a concept is specifically defined in one language, but not in another one. The concept is understood in more or less the same way in both languages, yet in one of them it is not labelled with any particular denomination.

Term discordance is influenced by various factors, including different methodologic perspectives in term creation and the contamination of these points of view in different languages. We must ask: when do we speak of related or synonymous terms? What is the identity of meanings? The dynamics of terminology must be perceived in relation to the endless modification of terminology systems required by scientific progress, as well as in relation to the internationalization of scientific communication.

### **3 The multilingual terminologic database and its structure**

The project *VEGA no. 1/0790/10 Contrastive Research of Grammar Metalanguage: Verb terminology (2010-2011)*, realized by the Department of Romance Studies and the Department of Slovak Language and Literature at the Faculty of Humanities of Matej Bel University, Banská Bystrica, aims to help improve international communication among linguists. The main task is to lead a contrastive study of linguistic terminology aiming at the creation of a multilingual database – including Slavic languages (Slovak) and Romance languages (French, Spanish, Italian) – of linguistic terms referring to the subject of ‘the verb’ (verbum).

The database of linguistic terms will proceed from an analysis of linguistic texts published in grammars, scientific monographs and papers, laying stress on the current state of research. The main goals of the project’s first stage are the following: to reflect on the content of the database according to the principles and methods of terminologic work, to establish the technical parameters of a computer platform and to create a conceptual structure of the subject area, linguistic terminology. The second stage of the project focuses on parallel analysis of linguistic texts, establishing equivalence between concepts and, consequently, terms in different languages, and the creation of a multilingual dictionary of terms. In the dictionary, definitions of individual terms will reveal discordances between the Slovak language and Romance languages (Spanish, French and Italian) in representations of terms. The project’s primary output will consist of a multilingual dictionary of linguistic terminology and a computer database.

#### **3.1 Terminologic database format**

The creation of terminologic files requires the application of common principles, as presented in international standards. The basic unit of terminologic sets is an entry containing all the information about a term, presented in a logical and accessible way.

With regard to ISO 10241, the document establishing the principles behind the creation of international terminologic standards, we distinguish two kinds of information in a terminologic entry: fundamental information (term, definition, source of definition, context, source of context, etc.) and complementary information (pronunci-

ation, graphic representation, etc.) Taking into consideration the particularities of linguistic terminology, a terminologic entry can contain other information, allowing one to identify the term (and the corresponding concept) in an unambiguous way. This information concerns, among other things, the theoretical approach which is applied in the definition of the term. In the case of linguistic terms, the terminologic entry's format is based on one widely used in the field, containing linguistic, encyclopaedic and administrative information.

The linguistic information related to the entry term consists of the following: the entry term in its basic form, the language of the entry term, the basic grammatical properties of the term, its etymologic origin, pronunciation, abbreviated form, synonyms, derived terms, collocations, a language note and its source, and the context of the term and its source.

The encyclopaedic information consists of the definition of the term and its source, the graphic representation of the term, its subject domain, subdomain and discipline, associated theoretical approach or school, a technical or encyclopaedic note and its source, hyperonyms, isonyms, hyponyms and antonyms that help pinpoint the position of the term in the system, its related terms, an equivalent of the term in another language, the degree of equivalence expressed by a percentage. In the case of a partial equivalence, the terminological entry contains an explicit explanation of the differences in meaning between the entry term and its foreign equivalent.

The terminologic entry contains administrative information consisting of the following: the code of the entry's author, the institution code, the research project code, the date of creation, the date of completion or most recent correction, the date of verification and finally the entry's identification number. Information can be added step by step, which is an advantage of using a computer database to compile data. By computerizing the processing of terms, we are able to present a great deal of information related to a single term. However, it is important to specify which information is the most important with regard to the aims and the character of the project.

### **3.2 The format of terminologic entries in the TermSlov database**

The TermSlov computer database has been created by the team of researchers involved in the VEGA project No. 1/0790/10 and it can be accessed through the computer network of the Faculty of Natural Sciences at Matej Bel University. It consists of a set of interrelated entries on verb terminology (see Figures 1 and 2).

All the information in the terminologic entry is related to the entry term. All of it is presented in the language of the entry term (except for the information related to equivalence, which appears in Slovak) or using the codes established and applied within VEGA project.

Each member of the research team has the power to create new terminologic entries. The database's basic functions are as follows:

- present detailed information on entry terms,
- connect terminologic entries in Slovak with entries in other languages by determining the equivalence of terms,

- connect each terminologic entry with entries of related terms in the same language,
- search for a terminologic entry by entering a term,
- make existing terminologic entries visible to all members of the team by entering an access code,
- print and save terminologic entries.

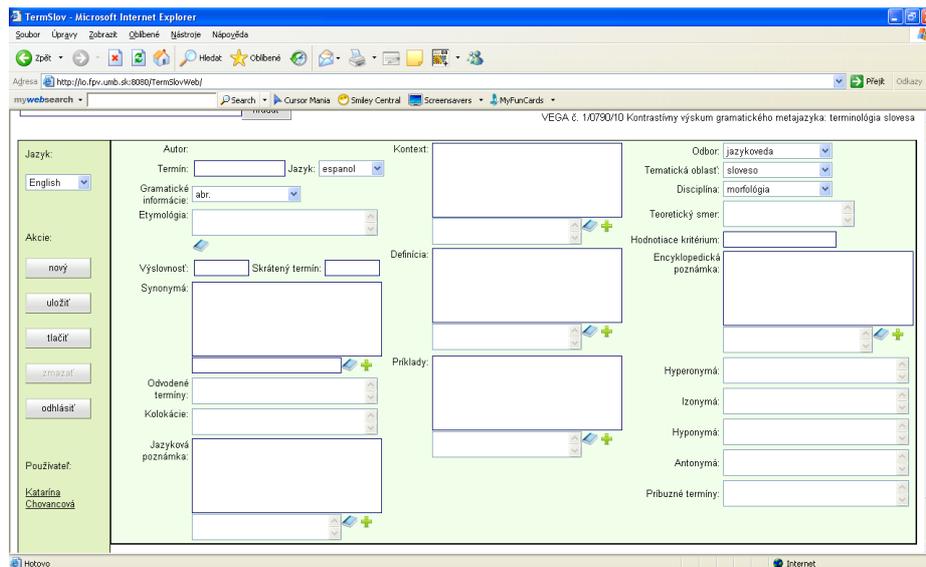


Fig.1. Screenshot of the TermSlov terminologic database interface

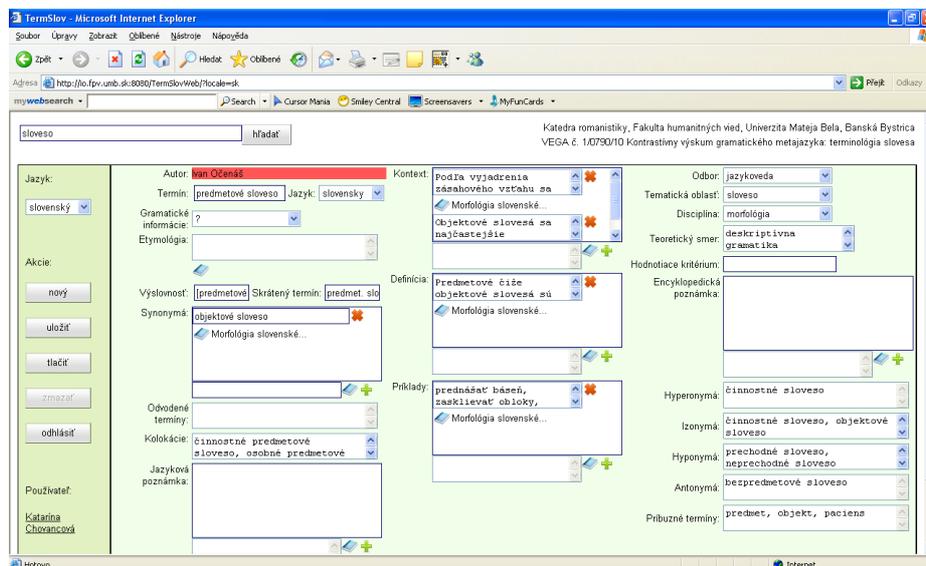


Fig.2. Example of a terminologic entry (*tranzitívne sloveso*) in the TermSlov database

The most important information stated in a terminologic entry is the entry term itself. It appears in its basic grammatical form (in Slovak, French, Spanish, or Italian, with an indication of the language it belongs to). Basic grammar information about the entry term is given in accordance with the lexicographic practice of the term's language and according to the format and abbreviations established by the research team. For example, the term *pomocné sloveso*'s word class as is given as *neutrum*. This is followed by etymological information on the entry term, with indication of the source, pronunciation of the term in the form of a phonetic transcription, as well as an abbreviation of the term, if applicable.

The next part displays synonyms of the entry term. These consist of other linguistic expressions of the entry term in the same language and must be confirmed. Up to five synonyms can be added. For example, *neplnovýznamový funktor* is listed as a synonym of *pomocné sloveso*. The subsequent items indicate derived terms, such as *slovesný* in case of the entry term *sloveso*, as well as collocations of the entry term. Collocations are sequences of words that often co-occur and include the entry term. The linguistic note specifies any linguistic particularities of the entry term, such as, for example, an irregular conjugation. In the terminologic entry of the French term *verbe auxiliaire* we read: 'Používa sa väčšinou v množnom čísle. Použitie tvaru auxiliaires je častejšie ako verbes auxiliaires.'<sup>5</sup> Each linguistic note has its source clearly indicated. Up to five linguistic notes can be added. The context field displays an original text attesting to the entry term. The context can be taken from various sources and it must contain the entry term. Let us take a closer look at the context for the term *pomocné sloveso*:

'Spojenia pomocných slovies s plnovýznamovými pomenovaniami sa hodnotia ako zložené vetné členy, napr. zložený prísudok, zložený vetný základ, zložený podmet, zložený predmet, zložený doplnok. S rozličnou vetnočlenskou platnosťou súvisí tvar pomocného slovesa, ktoré môže byť v určitom aj neurčitom tvare: začína pršať, začnite postupovať, začínajúc pomáhať, začína odporovať...' (Morfológia slovenského jazyka, 1966, p. 363)

and the context of the French term *verbe auxiliaire*:

'En général, lorsque plusieurs verbes coordonnés sont à un même temps composé et ont le même sujet, on exprime l'auxiliaire avec le premier

<sup>5</sup> Normally used in plural. *Auxiliaires* is used more often than *verbes auxiliaires*. (translated by the author)

<sup>6</sup> Composed forms containing auxiliary verbs and full verbs are considered as complex syntactic groups, e.g. complex predicate, complex phrase basis, complex subject, complex object, complex attribute. Varying syntactic function implies variation of the form of the auxiliary verb, which can thus appear in an indefinite form: začína pršať (it's starting to rain), začnite postupovať (you're starting to make some progress), začínajúc pomáhať (starting to help), začína si odporovať (starting to contradict himself).

verbe seulement, si les divers verbes se conjuguent avec le même auxiliaire – et surtout quand ces verbes forment corps.’ (Grevisse – Goose, 2005, p. 1221)<sup>7</sup>

The definition, one of the key items in the database, is chosen from an original source in the language of the entry term. Even if, in some cases, the definition can be proposed by a member of the research team or an independent researcher (thus an unpublished, orally formulated definition), priority is given to definitions taken from monographs, papers, encyclopaedias or dictionaries recognized by the scientific community. Here we cite a definition of *pomocné sloveso*:

‘Pomocné slovesá majú všetky gramatické náležitosti plnovýznamových sloviess – okrem dejového vecného významu. Spravidla stojí pri nich plnovýznamové pomenovanie, napr. sloveso v infinitíve, ktorým sa vyslovuje vecný obsah gramatické prísudku alebo vecného základu. Pomocné slovesá rozdeľujeme na fázové, limitné, modálne a sponové.’ (Morfológia slovenského jazyka, 1966, pp. 362-363)

The definition is followed by a ‘graphic representation of the entry term’. Here, the concept is expressed in the form of a diagram or another graphical representation.

In the field dedicated to examples, we see verbs of the given kind: *môžeme písať ceruzkou* (we can write in pencil), *ide ho zadusiť od hnevu* (he’s about to suffocate from anger), *začína sa oteplievať* (it is starting to get warmer), *stal sa veliteľom* (he became an officer), *už je chladno* (it’s getting cold) illustrate *pomocné slovesá*, while *être* (to be) and *avoir* (to have) are examples of *verbes auxiliaires*.

Some items identify the position of the entry term (and the concept it expresses) within the system. They refer to the scientific domain (linguistics), subject (verb terminology), linguistic discipline (morphology or syntax), and theoretical approach or school that uses the term. The encyclopaedic note relates particular features of the entry term from the conceptual (semantic) point of view and can be related to the theoretical approach or the evaluation criterion. The position of the entry term in the system is articulated through hyperonyms, isonyms, hyponyms, antonyms and related terms. In the case of *verbe auxiliaire*, the term *verbe* would be a hyperonym; *fázové sloveso*, *limitné sloveso*, *modálne sloveso*, *sponové sloveso* are hyponyms of *pomocné sloveso*; *verbe à sens plein* is an isonym<sup>8</sup> of *verbe auxiliaire*. Equivalents of the entry term in all the languages are also key information, linked to the previously created terminologic entries of equivalent terms.

<sup>7</sup> ‘In general, when multiple coordinated verbs in a composed form have the same subject and take the same auxiliary, the auxiliary appears with the first one only – et surtout quand ces verbes forment corps.’ (Grevisse – Goose, 1221)

<sup>8</sup> Isonyms are terms occupying a position on the same level of a terminologic system (Gouadec, 1990).

The 'Equivalent' field is related to that of 'degree of equivalence'. Equivalence of terms in various languages can be assessed as complete (90-100%) or partial (50-90%). 'Differences between equivalent terms' displays an explicit description of the meaning differences between a Slovak and a foreign term/concept. In cases of partial equivalence, 'differences' is obligatory information. Each equivalent bears the code of the researcher who proposed it.

The final part of the database contains administrative information, ranging from the name of the institution and the project to the name of the author and the verifier of the terminologic entry. It may look like this:

(INS) Faculty of Humanities, Matej Bel University, Banská Bystrica  
 (PRO) VEGA n. 1/0790/10 *Contrastive Research of Grammar Metalanguage: Verb Terminology*  
 (IDC) 00001/SLO/IOC<sup>9</sup>  
 (DAT) XX/XX/XXXX<sup>10</sup>  
 (AUT) IOC<sup>11</sup>  
 (DAT/KOR) XX/XX/XXXX<sup>12</sup>  
 (AUT/KOR)<sup>13</sup>  
 (DAT/OVR) XX/XX/XXXX<sup>14</sup>  
 (AUT/OVR)<sup>15</sup>

The structure of the terminologic database is not fixed, but flexible. It is possible to complete and modify it continually, according to the needs of the research team.

#### 4 Conclusion

Linguistic terminology has its particularities. While translating linguistic texts from Slovak to Romance languages (French, Italian, Spanish) and vice versa we encounter problems caused not only by the confrontation of individual linguistic systems, but also by different approaches applied in language research. A contrastive analysis of linguistic metaterminology aims to determine the degree of equivalence between terms in various languages. Creation of a terminologic tool – a parallel terminologic database – can contribute to eliminating the difficulties that arise in translation.

---

<sup>9</sup> Identification number of the terminologic entry in the following format: 00001/SLO/XXX (number/language code/code of the author of the terminologic entry – using conventions established by the research team)

<sup>10</sup> Date of the entry's creation

<sup>11</sup> Code of the entry's author

<sup>12</sup> Date of the entry's last modification or correction

<sup>13</sup> Code of the author of the entry's last modification or correction.

<sup>14</sup> Verification date

<sup>15</sup> Verifier code

## References

- [1] Čermák, F. (2001) *Jazyk a jazykověda*. Praha: Nakladatelství Karolinum.
- [2] Gouadec, D. (1990) *Terminologie. Constitution des données*. Paris: Afnor Gestion.
- [3] Grevisse, M. – Goosse, A. (2005) *Le bon usage – grammaire française*. De Boeck – Duculot.
- [4] Horecký, J. – Buzássyová, K. – Bosák, J. et al. (1989) *Dynamika slovných zásob súčasnej slovenčiny*. Bratislava : Veda.
- [5] Ibrahim, A. H. – Křečková, V. (2009) Terminologie linguistique : Quelques réflexions sur le traitement terminographique bilingue (slovaque-français). In Chovancová, K. – Klimová, K. – Reichwalderová, E. (ed.): *Studia Romanistica Beliana. Zborník príspevkov z medzinárodnej vedeckej konferencie "Románske štúdie: súčasný stav a perspektívy" 26. – 27. marca 2009 v Banskej Bystrici*. Banská Bystrica : UMB FHV. Pp. 151-160.
- [6] Janočková, N. (2005) Slovesá v lexikografickom spracovaní. In *Varia XII. Zborník príspevkov z XII. kolokvia mladých jazykovedcov v Modre – Piesku. 4. – 6. 12. 2002*. Zost. M. Šimková. Bratislava : Slovenská jazykovedná spoločnosť pri SAV.
- [7] Jurčáková, Z. (1999) Viacnásobné využitie terminologických údajov. In: *Knižničná a informačná veda XVIII. Zborník Filozofickej fakulty Univerzity Komenského*, Bratislava: UK, pp. 97-107.
- [8] Křečková, V. (2008) Slovensko-francúzska jazykovedná terminológia: metodológia terminologickej práce. In Ďuricová, A. (ed.) *Od textu k prekladu 3*, Praha: JTP, pp. 114-126.
- [9] Křečková, V. (2009) Slovensko-francúzska jazykovedná terminológia: identifikácia pojmov a modelovanie pojmových systémov. In Ďuricová, A. (ed.) *Od textu k prekladu IV*, Praha : JTP, pp. 73-83.
- [10] Lerat, P. (2010) *Variabilité et harmonisation terminologiques*. Atti Convegno Assiterm 2009. [online] In Publifarum, no. 12, 2010. [cit. 8.10.2010] Retrieved from: [http://publifarum.farum.it/ezine\\_articles.php?id=165](http://publifarum.farum.it/ezine_articles.php?id=165)
- [11] Levická, J. (2007) Slovenská terminologická databáza a jej úloha pri skvalitňovaní terminologických sústav. In *Slovenská terminologická sieť. Príspevky z konferencií*. [Online]. [Cit. 2010-03-17]. Retrieved from: [http://ec.europa.eu/dgs/translation/sts/conf\\_speech.html](http://ec.europa.eu/dgs/translation/sts/conf_speech.html)
- [12] Masár, I. (1991) *Príručka slovenskej terminológie*. Bratislava: Veda SAV.
- [13] Masár, I. (2000) *Ako pomenúvame v slovenčine. Kapitoly z terminologickej teórie a praxe*. Spisy SJS SAV 3/2000. Bratislava: SAV, 2000.
- [14] Mešková, L. (1995) Ponímanie frazeológie v súčasnej francúzskej lingvistike. In: *SLAVICA SLOVACA*, vol. 30, 1995, no. 1, pp. 56-66.
- [15] Navrátil, L. (2009) *Slovesá – kráľovský slovný druh*. Nitra : ENIGMA.
- [16] Nedobity, W. (1985). Konceptológia – náuka o pojmoch. In: *Kultúra slova*, vol. 19, no. 5, pp. 139-143.

- [17] Neveu, F. (ed.) (2006) *La terminologie linguistique: problèmes épistémologiques, conceptuels et traductionnels*. In *Syntaxe – Sémantique n° 7*. Caen : Presses Universitaires de Caen.
- [18] Normalizácia v oblasti terminológie, dokumentácie a informácií. In: *Centrum vedecko-technických informácií SR*. [Online]. [Cit. 2008-09-09]. Retrieved from: <http://www.cvtisr.sk/itlib/itlib014/norma.html>
- [19] Nižníková, J. – Sokolová, M. et al. 1998. *Valenčný slovník slovenských slovíes*. 1<sup>st</sup> ed. Filozofická fakulta Prešovskej univerzity v Prešove : SLOVACONTACT Prešov, 1998.
- [20] Ondrejovič, S. 1989. *Medzi slovesom a vetou. Problémy slovesnej konverzie*. Bratislava : VEDA, 1989.
- [21] Oravec, J. – Furdík, J. – Bajžíková, E. 1984. *Súčasný slovenský spisovný jazyk. Morfológia*. 1<sup>st</sup> ed. Bratislava : SPN, 1984.
- [22] Riegel, M. – Pellat, J.-CH. – Rioul, R. 1999. *Grammaire méthodique du français*. Paris : Presses Universitaires de France. 1999.
- [23] Slovenská terminologická sieť. In: *Slovenská terminologická sieť*. [Online]. [Cit. 2008-08-21]. Retrieved from: <http://www.slovenskaterminologickasiet.eu>
- [24] Sokolová, M. 1993. *Sémantika slovesa a slovesný rod*. Bratislava : VEDA, 1993.
- [25] STN ISO 10241. *Medzinárodné terminologické normy. Tvorba a úprava*. Slovenská technická norma, január 2000.
- [26] STN ISO 860. *Terminologická práca. Harmonizácia pojmov a termínov*. Slovenská technická norma, október 1998.
- [27] van Campenhoudt, M. 1996. *Abrégé de terminologie multilingue*. Centre de recherche Termisti, février 1996. [online] [cit. 30-09-2009] Retrieved from: <http://www.termisti.refer.org/theoweb1.htm>

# Bulgarian-Polish-Lithuanian Corpus – Recent Progress and Application

Ludmila Dimitrova<sup>1</sup>, Violetta Koseska-Toszewa<sup>2</sup>, Danuta Roszko<sup>2</sup> and Roman Roszko<sup>2</sup>

<sup>1</sup> Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia

<sup>2</sup> Institute of Slavic Studies, Polish Academy of Sciences, Warsaw

**Abstract.** The paper presents the recent progress of the first Bulgarian-Polish-Lithuanian experimental corpus and its possible applications in the language research. The corpus is currently under development in the framework of the joint research project “Semantics and Contrastive linguistics with the focus on a bilingual electronic dictionary” between IMI-BAS and ISS-PAS. The corpus contains more than three million words and comprises two corpora: parallel and comparable. Recent developments focus on 1) creation of a small aligned trilingual corpus and 2) an extension of the corpus.

## 1 Introduction

In recent decades many multilingual corpora were created in the field of corpus linguistics. The first Bulgarian-Polish-Lithuanian corpus is currently collected under the joint research project “Semantics and Contrastive linguistics with the focus on a bilingual electronic dictionary” between IMI-BAS and ISS-PAS for research purposes, (Dimitrova et al. 2009a). It is interesting to note that two Slavic languages are compared to a Baltic one – Lithuanian. Furthermore, the three languages are marginally present in the EU because of the later accession of these three countries to the EU.

The corpus contains more than 3 million words so far. All collected texts in the corpus are texts published in and distributed over the Internet, so copyright issues for the texts are not a concern. We note that a big problem arose when we started to compile the corpus due to the mismatch in proportion of translated texts in the three languages. It turned out that it is extremely difficult to find electronic texts of translations from Bulgarian to Lithuanian or vice versa. That’s why we assumed that the Polish language would build “a bridge” between Bulgarian and Lithuanian: for the Bulgarian-Polish and Polish-Lithuanian pairs of languages one can find freely available translations on the Internet. Our observations indicate that Polish literature is more frequently translated to Bulgarian or Lithuanian than Bulgarian or Lithuanian to Polish. However, the translated texts in the three languages must be of comparable size, otherwise the languages comprising the corpus would not be equally represented. Each language corpus requires a good selection of materials. We stress here that a three language parallel corpus cannot be a sum of three monolingual corpora. The multilingual parallel corpora consist of texts in one source language and texts in trans-

lation in all the other languages. Material in such corpora has to be organized so that each language is a source language. Furthermore, each of the languages could be investigated also through translations in other languages.

The parallel trilingual corpus comprises original texts in one of the three languages with translations in two others, texts of official documents of the European Union available through the Internet and texts, mainly fiction in other languages, translated into Bulgarian, Polish, and Lithuanian.

## 2 Structure and description of the corpus

The trilingual corpus comprises two corpora: parallel and comparable.

### 2.1 Parallel Bulgarian-Polish-Lithuanian corpus

The **Bulgarian-Polish-Lithuanian parallel corpus** contains more than 1 million words up to now. A part of the parallel corpus comprises original literary texts (fiction, novels, and short stories) in one of the three languages with translations in the other two, and texts of brochures of the European Commission, official documents of the European Union and the European Parliament, available through the Internet. The remaining part of the parallel corpus comprises texts, mainly fiction, in other languages translated into Bulgarian, Polish, and Lithuanian.

Thus, we classify texts as either

- **Core** – original literary texts (fiction, novels, and short stories) in one of the three languages with translations in the other two, aligned at the paragraph or sentence levels, for example: Stanisław Lem’s “Solaris” with Bulgarian and Lithuanian translations, Vladas Braziūnas’s “Yesterday is Tomorrow” with Bulgarian and Polish translations.
- **Translations** – texts in other languages translated into Bulgarian, Polish, and Lithuanian: literary works, texts of brochures and documents of the EC, EU, etc. Literary works include A. de Saint-Exupéry’s “Le Petit Prince”, Bulgakov’s “Master and Margarita”, Tolkien’s “The Lord of the Rings”, A. A. Milne’s “Winnie-the-Pooh”.

The recent result of our work is a small, aligned Bulgarian-Polish-Lithuanian corpus, which we briefly describe next.

A parallel text is a text placed alongside its translation or translations. Large collections of bi- or multilingual parallel texts are parallel corpora. An aligned corpus is a special kind of multilingual parallel corpus, that is the result of the process of parallel text alignment. The process of alignment aims at producing a set of corresponding sentences (original and its translation(s)) in both or more parts of the parallel text (one of the most well-known example of parallel text alignment is inscribed on the famous Rosetta Stone). In the end of the process that aligns two par-

allel texts, a merged document, called bi-text, composed of both source- and target-language versions of a given text that retains the original sentence order, is produced.

The alignment is a non-trivial task because of the role of the translator: some sentences can be split, merged, deleted, inserted or reordered during the translation. The software tools, generating bi-texts, are called alignment tools, or bi-text tools, which automatically align the original and translated versions of the same text. The tools generally match these two texts sentence by sentence. Collections of bi-texts form so called bi-text database(s).

An example of a multilingual aligned corpus is the MULTTEXT-East corpus for Orwell's 1984 (Dimitrova et al. 1998). Six CEE languages – Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovene – were involved in the first aligned version which was used in many applications. The MTE multilingual aligned corpus consists of six bi-texts: all six translations of the novel aligned with the English original. The alignment at the sentence level was performed automatically by means of some software packages and the obtained results were manually checked. The following examples (Fig. 1) illustrate some consecutive pairs of aligned sentences from the Bulgarian-English bi-text.

1-1 aligned sentences	<Obj.1.1.14.7>Потопи писалката в мастилото и за миг се поколеба. <Oen.1.1.15.7>He dipped the pen into the ink and then faltered for just a second.
1-2 aligned sentences	<Obj.1.1.14.8>Стомахът му се сви: началото беше съдбоносно. <Oen.1.1.15.8>A tremor had gone through his bowels.<Oen.1.1.15.9>To mark the paper was the decisive act.
1-2 aligned sentences	<Obj.1.1.23.16>Не беше много вероятно и въпреки това винаги, когато тя бе наоколо, той изпитваше странно чувство на неудобство, примесено със страх, дори враждебност. <Oen.1.1.24.16>That, it was true, was very unlikely.<Oen.1.1.24.17>Still, he continued to feel a peculiar uneasiness, which had fear mixed up in it as well as hostility, whenever she was anywhere near him.
2-1 aligned sentences	<Obj.1.1.29.5>Той седеше с изправен торс, мощният му гръден кош се издуваше и потрепваше, сякаш се готвеше да поеме удара на вълна.<Obj.1.1.29.6>Тъмнокосото момиче зад <b>Уинстън</b> крещеше: "Свиня!" <Oen.1.1.30.5>He was sitting very straight in his chair, his powerful chest swelling and quivering as though he were standing up to the assault of a wave.

Fig. 1

To align the texts of the parallel Bulgarian-Polish-Lithuanian corpus we used two language-independent freely-available software tools: MT2007 Memory Translation computer aided tool (TextAlign), and Bitext Aligner/Converter (Bitext2tmx aligner). TextAlign is available at <http://mt2007-cat.ru/index.html>. This software package segments and aligns corresponding translated sentences, contained in two rich text format

files. Bitext2tmx aligner (available at <http://bitext2tmx.sourceforge.net>) is a Java application. It works on any Java supported operating system (e.g. Windows, Linux, Mac OS X, Solaris), and is released under the GNU General Public License. Bitext2tmx aligner is a software tool that segments and aligns corresponding translated sentences, contained in two plain text files. These software packages have applications in computer-assisted translation. Both tools align bilingual texts without bilingual dictionaries, but the human editing is obligatory. The resulting aligned texts are similar.

At the first stage we used the align software tool to align the original text, for example Stanislaw Lem's Solaris in Polish, and its Bulgarian translation. At the second stage the procedure is repeated with the input pair being the original Polish text and its Lithuanian translation. At the third stage, after a comparison of the two output bitexts, Polish-Bulgarian and Polish-Lithuanian, we end up with a sequence of triples: a sentence in Polish and its translations in Bulgarian and Lithuanian.

The following example (Fig. 2) presents an excerpt from the aligned at the sentence level texts of Stanislaw Lem's Solaris (a part of the **Core corpus**) using TextAlign software:

```
<tu tuid="0000000006">
  <tuv xml:lang="polish">
    <seg>Podniósłszy oczy, zobaczyłem przez wypukłą szybę ściany studni i wyżej schyloną nad nią twarz Moddarda.</seg>
  </tuv>
  <tuv xml:lang="bulgarian">
    <seg>Вдигнах очи и през изпъкналото стъкло видях стените на кладенеца, а горе - надвесеното лице на Модард.</seg>
  </tuv>
  <tuv xml:lang="lithuanian">
    <seg>Pakėlęs akis, pro išgaubtą šulinio sienos langą pamačiau pasilenkusio Modardo veidą.</seg>
  </tuv>
</tu>
<tu tuid="0000000007">
  <tuv xml:lang="polish">
    <seg>Znikła zaraz i zapadła ciemność, bo z góry nałożono ciężki ochronny stożek.</seg>
  </tuv>
  <tuv xml:lang="bulgarian">
    <seg>Но то веднага изчезна и наоколо стана тъмно, защото отгоре поставиха тежкия предпазен конус.</seg>
  </tuv>
  <tuv xml:lang="lithuanian">
    <seg>Paskui veidas dingo ir pasidarė tamsu, nes viršuje kažkas uždėjo sunkų apsauginį konusą.</seg>
  </tuv>
</tu>
```

Fig. 2

The trilingual corpora aligned at sentence level are useful for many NLP applications: for instance, in systems for machine-aided human translation, or for training of programming tools for machine translation. In addition, they are prerequisite for many areas of linguistic research, and can also be used for retrieval of linguistic information or for producing concordances.

The next example (Fig. 3) presents a part of the parallel specialized texts of the trilingual corpus (**Translations corpus**) – texts of documents of the European Commission, the European Union, the Lisbon Treaty, 7FP of the European Commission documents, in Bulgarian, Polish, Lithuanian, and English as a “hub language”. The table shows aligned at paragraph-level texts from **ENWISE**<sup>1</sup> Expert Group Report “Executive Summary (30 January 2004)”.

<p><b>Bulgarian</b> <a href="http://ec.europa.eu/research/science-society/pdf/sc_soc_wasted_talents_bg.pdf">http://ec.europa.eu/research/science-society/pdf/sc_soc_wasted_talents_bg.pdf</a>  Преходният период води до реструктуриране на системата за научно-изследователска дейност в ENWISE-страните и най-общо се характеризира с рязък спад във финансирането на науката, със съкращаване на научния персонал и ликвидиране на изследванията, свързани с военната промишленост.</p>
<p><b>Polish</b> <a href="http://ec.europa.eu/research/science-society/pdf/sc_soc_wasted_talents_pl.pdf">http://ec.europa.eu/research/science-society/pdf/sc_soc_wasted_talents_pl.pdf</a>  Okres transformacji ustrojowej doprowadził do zmian w systemie badań naukowych w krajach Enwise. Gwałtowne zmniejszenie funduszy na finansowanie badań naukowych i, tym samym, zmniejszenie liczby pracowników naukowych oraz spadek znaczenia przemysłu zbrojeniowego i gałęzi z nim związanych, to najbardziej widoczne skutki wspomnianych przekształceń.</p>
<p><b>Lithuanian</b> <a href="http://ec.europa.eu/research/science-society/pdf/sc_soc_wasted_talents_lt.pdf">http://ec.europa.eu/research/science-society/pdf/sc_soc_wasted_talents_lt.pdf</a>  Pereinamajame laikotarpys mokslinių tyrimų sistemos Enwise šalyse buvo pertvarkomos. Tuo metu mokslui buvo skiriama vis mažiau lėšų, mažinamas mokslo darbuotojų skaičius, nyko karinė ir su ja susijusi pramonė.</p>
<p><b>English</b> <a href="http://ec.europa.eu/research/science-society/pdf/sc_soc_wasted_talents_en.pdf">http://ec.europa.eu/research/science-society/pdf/sc_soc_wasted_talents_en.pdf</a>  The transition period has led to the restructuring of the research systems in the Enwise countries and can generally be characterised by the sharp decline in funding allocated to science, the decrease of the research population and the disappearance of the military and associated industries.</p>

**Fig. 3**

<sup>1</sup> *Enlarge 'Women In Science' to East* project, **ENWISE** for short

## 2.2 Comparable corpus

The trilingual comparable corpus includes two type of texts: (1) literary works in Bulgarian, Polish and Lithuanian with the text sizes being comparable across the three languages, and (2) excerpts from electronic media newspapers (E-media texts) distributed via the Internet.

We have included literary works (novels or short stories) of some famous Bulgarian, Polish, and Lithuanian writers, among them novels of D. Dimov, D. Talev, J. Radichkov, P. Vezhinov – in Bulgarian; of St. Lem, R. Kapuściński, St. Žeromski – in Polish; of J. Žilinskas, I. Šeinius – in Lithuanian.

The E-media text is a description of the same event in the three languages. Therefore the three descriptions have the same thematic center and they should essentially present the same thematic content, essence). We specify such text as “parallel descriptions of content”. Obviously it cannot be expected that such texts be comparable in size unlike classical comparable corpora. The English text, describing the same event, is also included.

The E-media texts are stored as records with very simple structure:  
 <identification number> <URL of the Bulgarian web-page, Bulgarian text> <URL of the Polish web-page, Polish text> <URL of the Lithuanian web-page, Lithuanian text> <URL of the English web-page, English text>.

An example of parallel descriptions of content is shown in Fig. 4: E-media texts – record with identification number Bg-Pl-Lt-023. We note that the event “Germany qualified third at the 2010 FIFA World Cup championship having won against Uruguay 3:2” is described as follows: in Bulgarian by 253 words, in Polish – 286, in Lithuanian – 709, while one BBC description contains 708 words.

<b>Bulgarian</b> <a href="http://novini.dir.bg/news.php?id=6705372">http://novini.dir.bg/news.php?id=6705372</a>
<b>Германия спечели малкия финал, Форлан и Уругвай с поредни шедеври</b> <b>Sport1 - 10 Юли 23:27</b> Германия победи с 3:2 Уругвай в мача за третото място на Мондиал 2010, а срещата се првърна в поредния зрелищен и резултатен двубой с участието на южноамериканците, които отстъпиха със същия резултат и в 1/2-финала срещу Холандия. Съперниците си размениха на два пъти водачеството в срещата, но в крайна сметка макар и с резервите Бундестима спечели третото място, както и преди 4 години пред родна публика. Головете за победителите вкараха Мюлер, Янсен и Кедира, а за урусите точни бяха Кавани и Форлан. В 19-тата минута Муслера изби далечен удар, но право в краката на Мюлер, който не сгреша при добавката и реализира петия си гол за шампионата, откривайки резултата. В 28-та минута Кавани беше изведен добре между разредената отбрана на Германия и с премерен диагонален изстрел матира Бут и изравни. В 52-та минута звездата на южноамериканците Диего Форлан вкара най-красивия гол в мача, засичайки от воле центриране отдясно и така и той се изравни на върха на голмайсторската класация с 5 попадения. Само 4 минути по-късно Янсен беше оставен непокрит от Лугано и колегите му в отбраната на Уругвай и отблизо простреля с глава Муслера. В 82-та минута Сами Кедира се възползва от поредната грешка в защитата на урусите и подобно на Янсен отблизо с глава реализира и дари победата на своите. В последната трета минута на даденото от съдията продължение Форлан отново блесна, но шутът му от пряк свободен удар нацели гредата зад Бут и така Бундестима се поздрави с третото място.

<p><b>Polish</b>  <a href="http://mundial.wp.pl/title,Urugwaj-Niemcy-na-zywo,wid,12462530,wiadomosc.html">http://mundial.wp.pl/title,Urugwaj-Niemcy-na-zywo,wid,12462530,wiadomosc.html</a></p>
<p><b>Niemcy z brązowym medalem mistrzostw świata</b>  <b>Reprezentacja Niemiec wygrała z Urugwajem 3:2 i zdobyła brązowy medal XIX finałów piłkarskich mistrzostw świata. Niemcy powtórzyli wynik sprzed czterech lat, kiedy to wywalczyli brąz na własnym terenie.</b>  Pierwszą bramkę zdobyli Niemcy. W 18. minucie Bastian Schweinsteiger uderzył z 30 metrów, piłkę odbił Fernando Muslera, ale natychmiast doszedł do niej Thomas Mueller, którego dobitka w sytuacji sam na sam dała Niemcom prowadzenie. Urugwajczycy nie złożyli jednak broni.  10. minut później Perez odebrał piłkę w środku pola Schweinsteigerowi i podał do Suareza, który popędził z szybką kontrą. Po jego podaniu w pole karne Cavani strzałem w długi róg pokonał Hansa Joerga-Butta. Był więc remis i piłkarze Celestes zaczęli grać odważnie.  Na trzy minuty przed końcem pierwszej połowy znakomitą sytuację miał Suarez jednak jego strzał przeszedł tuż obok słupka.  Drugą połowę od ataków rozpoczęli Celestes. Dwie minuty po przerwie szansę miał ponownie Suarez ale jego strzał odbił Butt. Po chwili nie miał jednak nic do powiedzenia. Z prawej strony dośrodkował Alevaro a kapitalnym uderzeniem popisał się Diego Forlan i piłka wpadła do siatki. Bramkarz reprezentacji Niemiec nawet nie drgnął!  Urugwajczycy nie cieszyli się jednak długo z prowadzenia.  W 57. minucie fatalny błąd popełnił Fernando Muslera który minął się z piłką i wykorzystał to Jansen który głową trafił do pustej bramki.  W 82. minucie Niemcy zdobyli trzeciego gola. Po zamieszaniu w polu karnym gola głową zdobył Sami Khedira. Była to decydująca bramka tego meczu i to Niemcy zdobyli brązowy medal mistrzostw świata. Choć trzeba przyznać że Forlan w ostatniej akcji meczu był blisko pokonania Butta, ale jego uderzenie z rzutu wolnego trafiło w poprzeczkę. Celestes zajęli tylko 4 miejsce ale swoją grą i ambicją zyskali na pewno wielu sympatyków na całym świecie.</p>
<p><b>Lithuanian</b>  <a href="http://sportas.delfi.lt/worldcup2010/pasaulio-cempionato-bronzos-medaliai--ir-vel-vokietijos-rinktinei.d?id=34358887">http://sportas.delfi.lt/worldcup2010/pasaulio-cempionato-bronzos-medaliai--ir-vel-vokietijos-rinktinei.d?id=34358887</a></p>
<p><b>Pasaulio čempionato bronzos medaliai – ir vėl Vokietijos rinktinėi</b>  Atnaujinta 23.31 val., www.DELFI.lt 2010 liepos mėn. 10 d. 21:17  <b>Vokietijos futbolo rinktinė galėtų būti tituluojama pačia stabiliausia komanda planetoje. Pasaulio čempionate Pietų Afrikos Respublikoje vokiečiai pakartojo 2006 metų pirmenybių rezultatą ir užėmė trečiąją vietą. Rezultatyviame ir permainingame mūšyje dėl bronzos medalių šeštadienį Vokietijos futbolininkai Port Elizabeto mieste „Nelson Mandela Bay“ stadione, talpinančiame 48 tūkst. žiūrovų, 3:2 nugalėjo Urugvajaus nacionalinę komandą.</b>  18-ą min. vokiečius į priekį išvedė penktą kartą čempionate pasižymėjęs Thomasas Muelleris, bet 28-ą min. rezultata išlygino Edinsonas Cavani.  Situacija vertėsi aukštyn kojomis, kai 51-ą min. taip pat penktą savo įvartį įmušė Diego Forlanas ir jau Urugvajus atsidūrė priekyje. Bet urugvajiečių laimė buvo trumpa: jau 56-ąją Marcellis Jansenas atstatė pusiausvyrą, o 82-ą min. Vokietijos rinktinėi pergalę išplėšė Sami Khediros įvartis.  Vyriausiasis mačo arbitras – meksikietis Benito Archundia Tellezas.  Jeigu urugvajiečiai šiame mače žaidė stipriausios sudėties, tai Vokietijos komandos stovykloje nuotaikos prieš mačą buvo ne pačios geriausios. Dėl gripo ir traumų vyriausiasis ekipos treneris Joachimas Loewas į starto sudėtį neįtraukė vartininko Manuelio Neuerio, gynėjo Philippo Lahmo, saugo Lukaso Podolski ir puolėjo Miroslavo Klose. Nė vienas iš jų aikštėje taip ir nepasirodė.  <b>Rungtynės</b>  Susitikimas prasidėjo vokiečių atakomis, bet jų varžovai gynėsi sėkmingai. Netrukus jau Urugvajaus atakų lyderis Diego Forlanas turėjo keletą galimybių baudos smūgiais ir po vieno iš jų kamuolys skriejo šalia vartų.  10-ą min. fantastiška galimybė nepasinaudojo Vokietijos rinktinės gynėjas Arne Friedrichas, po kampinio galva iš kelių metrų nukreipęs kamuolį į vartų skersinį. Vokiečių spaudimas augo.  Ir savo tikslą Vokietijos komanda pasiekė 18-ą min. Ne pačioje pavojingiausioje atakoje Bastianas Schweinsteigeris smūgiu iš toli patikrino Urugvajaus vartininko Fernando Musleros budrumą. Šiam nepavyko sugauti kamuolio ir teko jį atmušti prieš save. Tuo metu vangi Urugvajaus gynyba pražiopsojo spurtavusį Thomasą Muellerį, kuris pirmasis suskubo prie kamuolio ir pasiuntė jį į vartus – 1:0.  25-ą min. šalia vartų atsidūręs D. Forlanas buvo arti tikslo, bet po jo smūgio galva kamuolį į užribį nukreipė gynėjas Peras Mertesackeris. Vokiečiai atsakė kontrišpuoliu, kurio smaigalyje atsidūręs Sami</p>

Khedira galėjo padvigubinti skirtumą. Vis dėlto F. Muslera išmušė kamuolį, o arbitras dar ir užfiksavo nuošalę.

Savo persvara Vokietija džiaužėsi neilgai. Diego Perezas aikštės centre perėmė kamuolį iš B. Schweinsteigerio ir išvedė tris savo komandos žaidėjus į greitą ataką prieš du gynėjus. Luisas Suarezas tiksliai perdavė kamuolį Edinsonui Cavani ir šis nestodamas išlygino rezultatą – 1:1.

Įsismarkavus lietuvių komandos ėmė žaisti ramiau, tačiau vėl ryškėjo Vokietijos komandos teritorinis pranašumas ir urugvajiečiams vis ilgiau teko gintis savo aikštės pusėje.

42-ą min. vokiečiai suklydo gynyboje ir L. Suarezas dešiniuoju kraštu išbėgo netrukdomas vartų link. Vokietijos komandą gelbėjo puikai poziciją užsiėmęs vartininkas Hansas-Joergas Buttas, privertęs varžovą kamuolį pasiūsti šalia virpsto.

Kėlinio pabaigoje abi komandos keitėsi greitomis atakomis, tačiau sužaidus 45 min. rezultatas išliko lygus.

48-ą min. puikiai žaidė vokiečių vartininkas H.J. Buttas, laiku išbėgęs ir atlaikęs E. Cavani smūgį, o iškart po to spėjęs pašokti bei atmušti ir pakartotinį L. Suarezo bandymą.

Vokietijos rinktinė patyrė netikėtą varžovų kirtį 51-ą min. Urugvajaus rinktinės puolėjas Diego Forlanas demonstravo puikią techniką, baudos aikštelės prieigose skersai mušęs kamuolį į žemę ir šis nenuspėjama trajektorija nuskriejo į vartus. Šį kartą H. J. Buttas nesitikėjo tokio varžovo veiksmo ir savo komandos išgelbėti nebesugebėjo – 2:1.

Tačiau Urugvajaus laimė buvo trumpa. Jau 56-ą min. vokiečių gynėjas Dennisas Aogo iš tolimo dešinio krašto pakėlė kamuolį į varžovų baudos aikštelę, kur aukščiau visų išsoko Marcellis Jansenas ir galva pasiuntė kamuolį į vartus – 2:2.

Rezultatyviame mače tempas nemažėjo. Abi komandos ir toliau aktyviai bei greitai veržėsi į atakas.

76-ą min. aikštėje tik pasirodęs Vokietijos rinktinės puolėjas Stefanas Kiesslingas techniškai apgavo gynėjus ir pavojingai iš maždaug 14 metrų smūgiavo į vartus. Šįkart vartininkas F. Muslera savo darbą atliko sėkmingai.

80-ą min. dėmesio centre vėl atsidūrė S. Kiesslingas. Šįkart puolėjas išsoko prieš vartus, bet nepasiekė kamuolio ir „neuždarė“ puikaus perdavimo.

82-ą min. Vokietijos komanda žengė didelį žingsnį pergalei link. Kilus sumaiščiui Urugvajaus baudos aikštelėje, Sami Khedira galva švelniai mestelėjo kamuolį per vartininką F. Muslerą ir šis dar kartą nesugebėjo išgelbėti savo ekipos – 3:2.

Paskutinėmis minutėmis vokiečiai gynyboje neužsidarė ir netrukus S. Kiesslingas vėl neišnaudojo puikios progos. Prieš vartininką iššokęs puolėjas spyrė kamuolį gerokai virš vartų.

Paskutinio savo šanso neišnaudojo D. Forlanas. Slekant paskutinėms teisėjo pridėto laiko sekundėms Urugvajaus atakų vedlys baudos smūgiu pasiuntė kamuolį į skersinį ir mačas baigėsi Vokietijos rinktinės pergale.

#### English

[http://news.bbc.co.uk/sport2/hi/football/world\\_cup\\_2010/matches/match\\_63/default.stm](http://news.bbc.co.uk/sport2/hi/football/world_cup_2010/matches/match_63/default.stm)

#### Uruguay 2-3 Germany

##### **Sami Khedira nodded in a late winner as Germany saw off the spirited challenge of Uruguay in Port Elizabeth to finish in third place for the second World Cup running.**

It seemed the three-time champions were set for a comfortable evening when Thomas Mueller capped a period of dominance with a straightforward finish after Fernando Muslera spilled a long-range drive from Bastian Schweinsteiger.

Uruguay forced their way back into proceedings through Edinson Cavani before Diego Forlan put them ahead with an acrobatic volley shortly after half-time.

But more poor goalkeeping from Muslera allowed Marcell Jansen to equalise and Khedira then capitalised on Uruguay's failure to clear a Mesut Ozil corner to ensure Germany would end their tournament on a happy note.

The third-place play-off is often billed as a match that no team wants to play - but this encounter made a mockery of any such suggestion.

Uruguay exceeded all expectations in reaching the last four and, with a parade through the streets of Montevideo and a date with President Jose Mujica planned for Monday, they were desperate to finish their campaign on a high.

Germany, meanwhile, are a hugely talented young squad who will be together for some years and viewed the meeting with La Celeste as another stepping stone in their development.

Oscar Tabarez's men signalled their intent by restoring fit-again captain Diego Lugano and star striker Luis Suarez, suspended for the last-four defeat by the Netherlands, to the starting line-up.

Although Germany made five changes from the side beaten by Spain, Philipp Lahm and Lukas Podolski were suffering from a virus, Miroslav Klose had a bad back and coach Joachim Loew was still able to

name an enviable line-up.  
 And it was clear from the opening exchanges that this was no friendly as Mueller saw a strike ruled out for handball, Dennis Aogo escaped with a caution for a terrible studs-up challenge on Diego Perez and Forlan curled a free-kick just over.  
 Germany, however, dominated the early possession and, after Arne Friedrich had nodded a Ozil corner against the crossbar, they opened the scoring.  
 Schweinsteiger unleashed a swerving drive from about 35 yards that was spilled by Muslera, offering Mueller a simple finish to draw level with David Villa and Wesley Sneijder on five goals.  
 Germany might have been expected to kick on from there but instead they sat back on their lead and invited Uruguay forward, while looking to launch the occasional counter-attack of their own.  
 The two-time champions duly settled into a tidy rhythm, showcasing their ability to turn defence into attack at frightening pace, and were soon back on level terms.  
 Schweinsteiger, captaining Germany in the absence of Lahm, was robbed of possession by Perez before Forlan fed Cavani and the 23-year-old Palermo striker calmly slotted past Hans Joerg Butt.  
 Despite persistent rainfall at the Nelson Mandela Bay Stadium, Uruguay were passing and moving beautifully and they should have gone ahead before the break when Forlan released Suarez, only for the Ajax frontman to drag a shot wide.  
 Germany, making a record fifth appearance in the third-place play-off, seemed to have lost their earlier intensity and that allowed the South Americans, full of desire and no shortage of quality, to pick up where they left off in the second period.  
 Forlan, Suarez and Cavani caused their opponents all manner of problems and, after Cavani and Suarez were denied by Butt, Forlan met Egidio Arevalo's cross with a superb volley to notch his fifth goal of the finals.  
 Uruguay deserved their advantage and, with Germany showing little potency in the final third, they looked set fair for a fine victory.  
 But Muslera's unconvincing attempt at claiming a Jerome Boateng cross allowed Jansen to divert the ball into an unguarded net and suddenly the momentum had shifted.  
 Uruguay, who have still not beaten a European team at the World Cup since 1970, rapidly began to tire and Germany substitute Stefan Kiessling twice squandered opportunities to grab the winner before Khedira made no mistake.  
 Tabarez's players refused to accept defeat and almost forced extra time, Forlan curled a 92nd-minute free-kick on to the bar.  
 But Germany held on and their celebrations at full-time showed exactly what this result meant to them.

Fig. 4

### 3 Applications in contrastive studies

In this paper we continue to comment on the applications of the trilingual corpus in contrastive linguistics studies. Other applications were described in Dimitrova et al. 2010. We present two examples – translations into Bulgarian, Polish, and Lithuanian from two different languages, one belonging to the Romance language group – Antoine de Saint-Exupéry's "Le Petit Prince" (original in French, an analytic language), and another belonging to the East-Slavic language group – Bulgakov's "Master and Margarita" (original in Russian, a synthetic language). This fact automatically poses a series of questions important both for the comparative grammar, but also the bilingual dictionaries in the three languages – Bulgarian, Polish, and Lithuanian. One has to stress the fact that a translation from a synthetic language, such as Russian creates one set of problems for Bulgarian, Polish, and Lithuanian, while one from French, which is an analytic language, poses another set of problems for these three languages.

An excerpt from “Le Petit Prince” (with the texts aligned at the paragraph level) is shown in the table in Fig. 5.

Bulgarian	Polish	Lithuanian
Показах прекрасното си произведение на възрастните и ги попитах дали рисунката им вдъхва страх.	Pokazałem moje dzieło dorosłym i spytałem, czy ich przeraża.	Savo kūrinį aš parodžiau suaugusiems ir paklausiau, ar jiems mano piešinio baisu.
Те ми отговориха: "Защо една шапка да вдъхва страх?"	Dlaczego kapelusz miałby przerażać? - odpowiedzieli dorośli.	Jie man atsakė: "O ko čia bus baisu skrybėlės?"
Моята рисунка не изобразяваше шапка. Тя изобразяваше змия боа, която смила слон. Тогава нарисувах вътрешността на змията боа, та да могат възрастните да разберат. Те винаги имат нужда от обяснения. Моята рисунка номер 2 беше такава:	Mój obrazek nie przedstawiał kapelusza. To był wąż boa, który trawił słońca. Narysowałem następnie przekrój węża, aby dorośli mogli zrozumieć. Im zawsze trzeba tłumaczyć. Mój rysunek numer 2 wyglądał następująco:	Aš buvau nupiešęs visai ne skrybėlę. Buvau nupiešęs smauglį, virškinantį dramblių. Tada nupiešiau smauglio vidų, kad suaugusiems būtų suprantamiau. Jiems visada reikia aiškinti. Mano piešinys numeris 2 buvo šitoks:

Fig. 5

Here we take a slight digression and focus a little on the translation of the title of a book whose text (in translation from French) is included in the corpus: the translations in Bulgarian, Polish and Lithuanian of Antoine de Saint-Exupéry “**Le Petit Prince**”. The Bulgarian title retains the definite article as in the French title – “**Малкият принц**”. Analogously, the pronouns’ declension is observed in Lithuanian: **Mažasis princas** (compared **mažasis**). This is not the case with the Polish title of Exupéry’s book, translated as “**Mały Książę**”. This title itself points us to Ajdukiewicz’s theory of incompletely articulated quantification (Ajdukiewicz 1974), which will discuss a little further on, when we talk about translating materials from Russian to Bulgarian, Polish and Lithuanian.

The table in Fig. 6 presents the aligned at the paragraph-level texts from Michael Bulgakov’s “Master and Margarita”.

Bulgarian translation	Polish translation	Lithuanian translation
<p><b>Първият</b> беше <b>самият</b> Михаил Александрович Берлиоз, председател на управителния съвет на едно от най-големите московски литературни обединения, наричано съкратено МАССОЛИТ, и редактор на реномирано литературно списание, а неговият млад спътник - поетът Иван Николаевич Понириов, известен с псевдонима Бездомни.</p>	<p><b>Ten pierwszy</b> był <b>to</b> Michał Aleksandrowicz Berlioz we własnej osobie, redaktor miesięcznika literackiego i prezes zarządu jednego z największych stowarzyszeń literackich Moskwy, w skrócie Massolitu, towarzyszył mu zaś poeta Iwan Nikołajewicz Ponyriow, drukujący się pod pseudonimem <b>Bezdomny</b>.</p>	<p><b>Pirmasis</b> buvo ne kas kitas, kaip Michailas Aleksandrovičius Berliozas, storo literatūros žurnalo redaktorius ir vienos iš stambiausių Maskvos literatūrinių asociacijų, sutrumpintai vadinamos MASSOLIT'u, valdybos pirmininkas, o jaunasis jo palydovas – poetas Ivanas Nikolajevičius Ponyrevas, pasirašinėjantis Benamio slapyvardžiu.</p>
<p>Като се озоваха под сянката на едва раззеленилите се липи, <b>писателите</b> се втурнаха най-напред към шареното павилионче с надпис "Бира и безалкохолни напитки".</p>	<p>Kiedy <b>pisarze</b> znaleźli się w cieniu lip, które zaczynały się już zazieleniać, natychmiast ostro ruszyli ku jaskrawo pomalowanej budce z napisem "Piwo i napoje chłodzące".</p>	<p><b>Pasiekę</b> vos sužaliavusių liepų pavėsi, <b>rašytojai</b> pirmiausia puolė prie margai dažytos būdelės su užrašu „Alus ir vanduo“.</p>
<p>Да, трябва да отбележим първото странно нещо в тази страшна майска привечер. Не само край павилиончето, но и по цялата алея, успоредна на улица "Малая Бронная", нямаше жива душа. В този час, когато човек сякаш нямаше вече сили да диша, когато слънцето, нажежило Москва, потъваше в суха мъгла някъде отвъд Садовое кольцо, никой не беше излязъл под липите, никой не беше седнал на <b>пейка</b>, аляята беше пуста.</p>	<p>Tu musimy odnotować pierwszą osobliwość tego straszliwego majowego wieczoru. Nie tylko nikogo nie było koło budki, ale i w równoległej do Małej Bronnej alei nie widać było żywego ducha. Choć wydawało się, że nie ma już czym oddychać, choć słońce rozprażywszy Moskwę zapadało w gorącym suchym pyłe gdzieś za Sadowoje Kolco – nikt nie przyszedł pod lipy, nikogo nie było na <b>ławkach</b>, aleja była pusta.</p>	<p>Taigi reikia paminėti pirmą šio baisaus gegužės vakaro keistenybę. Ne tik palei būdelę, bet ir visoje alėjoje, nusidriekusiėje lygiagrečiai su Malaja Bronaja gatve, nebuvo nė vieno žmogaus. Tuo metu, kai rodės, nebėr kuo nė kvėpuoti, kai saulė, įkaitinusi Maskvą, smigo pro sausas miglas kažkur už Sodų žiedo – niekas nesiglaudė po liepomis, niekas nesėdėjo ant suoloelių, alėja buvo tuščia.</p>

Fig. 6

We will comment now on some of the significant differences between the three languages which can be illustrated by examples of texts from the aligned corpus, namely translations from Russian in Bulgarian, Polish and Lithuanian. In translations from Russian – the first and foremost question is how to carry on the meaning of the definite article and the some forms of the past tense. In Polish and Russian, which do not have a definite article, we face Ajdukiewicz's theory of incompletely articulated quantification (Ajdukiewicz 1974).

Here the language quantifications are missing and it is not clear if one is talking about a concrete object, about any such object or about all objects mentioned in the sentence. The incompletely articulated quantification is not a language error. Polish and Bulgarian differ from the point of view of this language phenomenon. This is explained with differences in the morphological systems of the two languages. In Polish, this is a frequently observed phenomenon. In both languages the incompletely articulated uniqueness (definiteness) decreases the incompletely articulated quantification; and is rarity in Bulgarian, because the definite noun group possesses a definite article (Koseska, Gargov 1990: 138–139, Koseska 2006). In Lithuanian, an intermediate state is observed. In comparison with Polish, the phenomenon *incompletely articulated uniqueness of the noun group* appears rarely in comparison with Bulgarian, where such phenomena appear more frequently.

All of this is related to the formal restrictions in the formation of definite (called pronoun) forms in Lithuanian (see below). It is necessary to stress, however, that the *incompletely articulated quantification of the noun group* disappears as the result of a union of such a group with a verbal group, which determines the quantification meaning and removes the incompletely articulated quantification.

Let us compare with the following example from the corpus (Fig. 6):

<p>Като се озоваха под сянката на едва раззеленилите се липи, <b>писателите</b> се втурнаха най-напред към шареното павилионче с надпис "Бира и безалкохолни напитки".</p>	<p>Kiedy <b>pisarze</b> znaleźli się w cieniu lip, które zaczynały się już zazieleniać, natychmiast ostro ruszyli ku jaskrawo pomalowanej budce z napisem "Piwo i napoje chłodzące".</p>	<p><b>Pasiekę</b> vos sužaliavusių liepų pavėsi, <b>rašytojai</b> pirmiausia puolė prie margai dažytos būdelės su užrašu „Alus ir vanduo“.</p>
--	--	--

It is not clear in the Polish translation if one is talking about exactly these writers or some other writers. (Here we have an “incompletely articulated quantification” of the phrase *pisarze* //the writers//, which does not exist as a phenomenon in Bulgarian. In the example, „*писателите*“ //the writers// is a plural definite noun with the article –*me* //the//, that expresses uniqueness (definiteness) of the given set).

In Lithuanian (unlike Bulgarian) the form is *rašytojai* //writers// (a definite article is missing), the use of a definite article in Lithuanian is limited only to certain morphological forms: qualitative adjectives, some pronouns, particles, and numerals.

However, in contrast with Polish, the “*incompletely articulated quantification*” phenomenon is missing, as in this Lithuanian sentence one uses a complex construction with the participle *pasiekę* (infinitive form *pasiekti* //to get in(to)//. The single occurrence past tense participle (short form, masculine, plural) is combined with the form *rašytojai* //the writers//. In this case it is the phenomenon “*completely articulated quantification of the noun phrase*” that is observed.

Let us take a look at the example above (Fig. 6), in particular the Polish sentence, in which we observe a phenomenon known as a *tendency towards analytism*:

<b>Първият</b> беше <i>самият</i> Михаил Александрович Берлиоз, ...	<b>Ten pierwszy</b> był <b>to</b> Michał Aleksandrowicz Berlioz <i>we własnej</i> <i>osobie</i> .	<b>Pirmasis</b> buvo <i>ne kas kitas, kaip</i> Michailas Aleksandrovičius Berliozas,
---	---	--

In the Bulgarian sentence „**Първият** //the first// (with definite article *ят*) беше *самият* Михаил Александрович Берлиоз, ...“ is translated in Polish as „**Ten pierwszy** był **to** Michał Aleksandrowicz Berlioz ...“. There are two demonstrative pronouns in the noun group in the Polish sentence: **ten** and **to**, which is a signal for process to *analytism* in Polish language (Sosnowski 2011).

In Lithuanian one observes an analogous phenomenon (as in Bulgarian), see **първият** and **pirmasis** – both forms are defined with a definite article. The comparison of Bulgarian, Polish, and Lithuanian in relation to the studied phrases: *самият – we własnej osobie – ne kas kitas, kaip* //He itself// is further complicated by the fact that in the three languages there are different means used (depending on the different constraints in the three languages).

#### 4 Conclusion and future work

Our materials demonstrate well the connection between semantics and language confrontation in linguistics studies, which is impossible through simple “mechanical summation” of monolingual corpora. This is the advantage of parallel corpora in comparison with monolingual corpora.

Parallel corpora are the most effective means for the creation of bi- and multilingual dictionaries and contrastive grammars. This is of great importance not only for language confrontation, but also for the typology of the studied languages. One has to remember that parallel corpora comprise direct material for the evaluation of translations and their analysis will bring out the improvement of the quality of both traditional, human translation, and machine translation. Besides, texts extracted from parallel or aligned corpora prove the necessity of evaluating translations: it is common that in translation words get omitted or word meanings get changed.

The parallel and aligned corpora are the best resource for the development of bi- and multilingual lexical databases and different kinds of digital dictionaries. There one could find and extract many examples of the usage of the words from the corpus in a wide context.

The parallel and aligned corpora are successfully used as language materials for the training of translators, as well as in education – for language learning in schools and universities. That is why online free-use parallel texts are also useful educational resource.

The semantics annotation of our trilingual corpus is an interesting research problem but is not an easy task. A uniform annotation system for Bulgarian, Polish, and Lithuanian does not currently exist. The problems of the morphosyntactic annotation of the corpus have been discussed in (Dimitrova et al. 2009b).

We envisage an extension of the parallel and aligned corpus (in volume and through additional annotated information). Another future task is the Web-presentation of the corpus with a multilingual (in Bulgarian, Polish, Lithuanian, and also in English) interface.

## References

- [1] Ajdukiewicz K. (1974). *Logika pragmatyczna*. Warszawa. (In Polish)
- [2] Dimitrova, L., Koseska-Toszewa, V., Roszko, D., Roszko, R. (2010). Application of Multilingual Corpus in Contrastive Studies (on the example of the Bulgarian-Polish-Lithuanian Parallel Corpus). *Cognitive Studies/Études Cognitives*, vol. 10, SOW, Warsaw, pages 217-240.
- [3] Dimitrova, L., Koseska, V., Roszko, D., Roszko, R. (2009a). Bulgarian-Polish-Lithuanian Corpus – Current Development. *Proceedings of the International Workshop “Multilingual resources, technologies and evaluation for Central and Eastern European languages” in conjunction with International Conference RANPL’2009. Borovec, Bulgaria, 17 September 2009*, pages 1-8.
- [4] Dimitrova, L., Koseska, V., Roszko, D., Roszko, R. (2009b). Bulgarian-Polish-Lithuanian Corpus – Problems of Development and Annotation. In T. Erjavec (Ed.), *Research Infrastructure for Digital Lexicography*, pages 72-86, Ljubljana, 2009.
- [5] Dimitrova, L., Erjavec, T., Ide, N., Kaalep, H., Petkevic, V., Tufis, D. (1998). Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. *Proceedings of COLING-ACL '98: 17<sup>th</sup> International Conference on Computational Linguistics and 36<sup>th</sup> Annual Meeting of the Association for Computational Linguistics, Montréal, Québec, Canada*, Ed. by Christian Boitet & Pete Whitelock, pages 315-319. San Francisco, Calif.: Morgan Kaufmann.
- [6] Koseska-Toszewa V., Gargov G. (1990). *Bulgarian-Polish contrastive grammar*, vol. 2: Semantics category definiteness/indefiniteness. Sofia. (In Bulgarian)
- [7] Koseska-Toszewa V. (2006). *Gramatyka konfrontatywna bułgarsko-polska*, vol. 7: Semantyczna kategoria czasu. SOW, Warszawa. (In Polish)
- [8] Sosnowski W. (2011). Analytic Tendencies in Modern Polish and Russian. *Cognitive Studies/Études Cognitives*, vol. 11, SOW, Warszawa, pages 97-108.

# Bulgarian–Slovak Parallel Corpus

Ludmila Dimitrova<sup>1</sup> and Radovan Garabík<sup>2</sup>

<sup>1</sup> Institute of Mathematics and Informatics, Bulgarian Academy of Sciences  
ludmila@cc.bas.bg

<sup>2</sup> E. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava  
garabik@kassiopeia.juls.savba.sk

**Abstract.** The paper presents a currently collected parallel sentence-aligned Bulgarian-Slovak/Slovak-Bulgarian corpus, developed in the framework of the joint research project between IMI–BAS and LŠIL–SAS. The texts in the ongoing version of the corpus are automatically aligned at the sentence level. The whole corpus is oriented towards emphasizing the applicability of the digital bilingual data for computerized natural language processing, but also as a source of human readable information. The paper describes the structure of the corpus. Other existing bilingual resources for Slovak and Bulgarian in combination with other languages are briefly mentioned as well.

## 1 Introduction

For many so called low profile languages, there are no multilingual or bilingual resources readily available for the scientific community. Multilingual parallel corpora are a basic resource for contrastive and terminology studies, for research and development of machine and human translation systems, language analysis, automatic term extraction, semantic analysis, supervised and unsupervised NLP tools training, etc. The parallel sentence-aligned Bulgarian-Slovak/Slovak-Bulgarian corpus is currently developed in the framework of the joint research project between IMI–BAS and LŠIL–SAS, coordinated by L. Dimitrova and R. Garabík.

## 2 Short comparison of the Bulgarian and Slovak languages

We will mention here only the differences that result from different orthography traditions, because we are primarily dealing with the written language, where the orthography forms an inseparable part of language analysis. The most visible (but from linguistic point of view rather superficial) difference is the writing system – Bulgarian uses the Cyrillic alphabet, Slovak – Latin with additional letters with diacritical marks.

There are some features specific for both languages, which have an influence on the language grammar repeatedly in several different phenomena, and which we will summarise here.

- First significant feature is the analytic character of Bulgarian, and the synthetic character of Slovak.
- In the process of evolution of Bulgarian from a synthetic, inflectional language, to an analytic language, case forms were replaced by combinations of different prepositions with a common case form. Bulgarian has lost most of the traditional Slavic case system.

- Bulgarian exhibits several linguistic innovations in comparison to the other Slavic languages, namely a rich system of verbal forms, and has a grammatical structure closer to English or the Neo-Latin languages than Slovak. The definite article is one of the most important grammatical characteristics of the Bulgarian language which sets it off the rest of the Slavic languages. The definite article is a morphological indicator of the grammatical category of determination (definiteness). It is not just a simple suffix, nor is it a particle, but a word-forming morpheme, which is placed at the end of words in order to express definiteness, knowledgeability, familiarity, conversance. In Bulgarian, nouns, adjectives, numerals, and full-forms of the possessive pronouns and participles can acquire an article. Slovak, as most of the other Slavic languages, lacks the definiteness altogether.
- For Slovak, a special feature of masculine nouns exists, the animate feature of masculine gender. This feature is variously analysed either as a subattribute of masculine gender where the words belonging to the masculine gender split into two categories, or the Slovak is sometimes said to have two different masculine genders, the animate and inanimate one. For Bulgarian there is no animate category at all.

### 3 Standards and models for corpora encoding

In our work on the corpus we used our own experience in the area of construction of language resources and technologies and the widely accepted international standards and models [IBR00]. We used some of the basic widespread technologies and software packages for annotation and usage of digital bilingual resources.

The problem of reusability of the digital language resources has been discussed so many times. It is well known that the development of such resources is a sustained process that is frequently repeated again and again. Some language resources are not available in a common standard usable format or their distribution is hindered due to intellectual property rights, diverse and poorly documented encodings and other reasons. We tried to use open standards in our corpus compilation process and internal data formats and to use simple textual representation of the data wherever possible.

The multilingual corpora are parallel and comparable. A parallel corpus is defined as a corpus comprising source texts and their translations. The parallel corpora (similar to the dictionaries) can be bilingual or multilingual. A comparable corpus is defined as a corpus comprising not-identical, but similar in their content texts in one or more languages. It can be defined as a set of texts, collected independently from their languages and connected on the basis of their content, domain and communicative function.

#### 3.1 Bulgarian parallel corpora

The first multilingual parallel corpus with Bulgarian has been produced as a part of the MULTEXT-East corpus. It consists of George Orwell's *1984* in English and its translations into six Central and East European languages (CEE): Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovenian [DEI<sup>+</sup>98], with Slovak as a later addition. Recently, bilingual Bulgarian-Polish [DK09] and trilingual Bulgarian-Polish-Lithuanian [DKRR10] corpora have been constructed.

### 3.2 Slovak parallel corpora

The first bilingual parallel corpus has been the Slovak-Russian parallel corpus, produced as a common project of LŠIL and Saint-Petersburg State University [R. 06], accessible through a web interface. It was later followed by the parallel Slovak-French corpus [VG07], using very similar structure and interface. Both of these corpora are lemmatised and POS tagged.

The Slovak-Czech parallel corpus contains mostly translations of fiction between Czech and Slovak (in both directions), with a small part of independent translations from other languages (75 % English, the rest German, Polish, Italian, French and Ancient Greek) into both Czech and Slovak. Both Czech and Slovak parts of the corpus are morphologically analysed and lemmatised with the *morče* software, the Czech part uses the tagset used in the Czech National Corpus.

The Slovak-English corpus consists of original English language fiction and their Slovak translations. The texts were obtained from various sources; the bulk of the Slovak translations were already collected in the Slovak National Corpus, some of them were scanned, OCRed and then proofread. A small amount of OCR-induced errors remained in the texts, but presumably, it will not have significant effects on the overall corpus quality. The English texts are lemmatized and POS-tagged, using the TreeTagger software [Sch97].

All the parallel corpora use the Hunalign software [VNH<sup>+</sup>05].

## 4 Structure of the Bulgarian-Slovak/Slovak-Bulgarian corpus

The corpus currently contains translations of fiction in both languages, either from Slovak into Bulgarian or from Bulgarian into Slovak. The main part of the parallel corpus contains texts in other languages translated into both Bulgarian and Slovak.

The parallel corpus consists of two subcorpora: direct and translated.

- The direct Bulgarian–Slovak parallel subcorpus consists of original texts in Bulgarian, such as novels and short stories by Bulgarian writers and their translations in Slovak, and original texts in Slovak, such as literary works by Slovak writers and their translations in Bulgarian.
- The translated Bulgarian–Slovak parallel subcorpus consists of Bulgarian and Slovak translations of literary works in the third language.

The Bulgarian–Slovak corpus contains parallel texts, aligned at the sentence level. It is composed of translations of literary works (mainly fiction) in both languages, either from Slovak into Bulgarian or from Bulgarian into Slovak (or translations into both of these languages from a third language).

To align the text on the sentence level, we use the Hunalign software. The program foresees the use of a corresponding bilingual dictionary to ensure a higher accuracy of the alignment; however, no such dictionary has been available for the use with the corpus.

The first version of the corpus contains 376 200 words in parallel texts, aligned at the paragraph level and at the sentence level. The set of aligned texts includes Bulgarian novels: Dimitar Dimov's *Осъдени Души* (Doomed Souls) and Pavel Vezhinov's *Барьерата* (The Barrier) and their Slovak translations (*Odsúdené duše* and *Bariéra*), the novel of

Slovak writer Klára Jarunková *Brat mlčanlivého vlka* (The silent wolf's brother) and its Bulgarian translation (*Братът на мълчаливия вълк*), and the Slovak and Bulgarian translations of Jaroslav Hašek's *The Good Soldier Švejk*.

## 5 Format of the texts and search

---

```

1 <doc lang="sk" origlang="sk"
2     biblio="Odsúdené duše,
3         Slovenský spisovateľ,
4         Bratislava 1976,
5         preklad Milan Topoľský">
6 <s id="1" link="1">
7 Odsúdené      odsúdený      Gtfp1x 12
8 duše         duša          Ssfp1  04
9 </s>
10 <s id="2" link="2">
11 Dimitār      dimitār       %       01
12 Dimov        dimov         %       24
13 </s>
14 <s id="3" link="3">
15 Prvá         prvý          Nafs1  02
16 časť        časť          Ssfs1  03
17 Koniec       koniec        Ssis1  04
18 jedného     jeden         Nfns2  04
19 dobrodružstva dobrodružstvo Ssns2  01
20 </s>

```

---

**Fig. 1.** Example of Slovak vertical file, wordforms with lemmas, morphosyntactic tags and number of disambiguation possibilities, sentences with links into Bulgarian text.

The structural annotation allows the texts in the two languages (Bulgarian/Slovak and vice versa) to be aligned at the corresponding level in order to produce aligned bilingual corpora. Alignment means the process of relating pairs of words, phrases, sentences or paragraphs in the texts in different languages which are translation equivalent. Currently, the corpus is automatically aligned at the sentence level without the help of a bilingual dictionary. In order to achieve reasonable quality of the corpus, the alignment should be as precise as possible. The project aims to create a small (several thousand words) dictionary suitable for automatic alignment (i.e. the dictionary will not be meant to be a general dictionary for human consumption).

The access to the recent version of the corpus is provided via a simple web interface (Fig. 2).

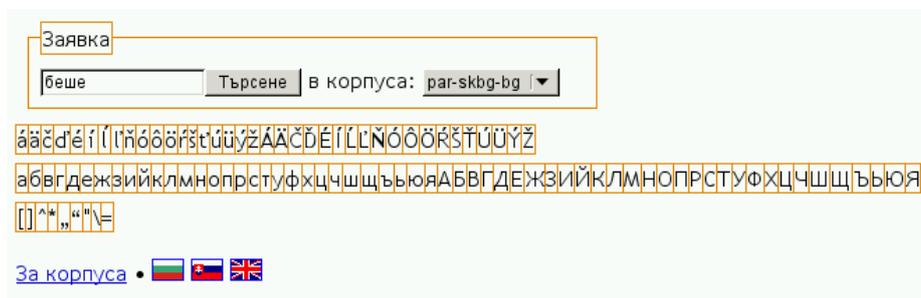


Fig. 2. Web search interface – a dialogue box in Bulgarian. A virtual Slovak/Bulgarian keyboard is visible under the query input field.

## 6 Applications of the bilingual corpus and future development

One of the uses of parallel corpora is the machine translation technology, either for training statistical models or translation evaluation (the current version of the corpus is well usable for the latter, but its size should be significantly increased before it will be useful for the machine translation training). The parallel and comparable corpora can be also used as a translation memory and language learning materials for training of (human) translators.

Besides, the bilingual corpora are envisaged to be used in education, in schools as well as universities for the foreign language instruction.

25852	Toto individuom ho istotne chce pokutovať za nedovolenú rýchlosť . Jack zvyčajne v takýchto prípadoch dupol na plyn , ale teraz mu akýsi dobrý duch pošepol , aby zavčasu zabrzdil . Keď zastal , všetci videli , že pred vozom ziva priepať po moste , ktorý odvílekol priväl .	Този индивид навярно искаше да го глоби за неправомерна скорост . В такива случаи Джек обикновено засилваше бързината , но сега някакъв <b>добър</b> дух го накара да натисне спирачките навреме . Когато спря , всички видяха , че пред колата зееше пропастта на един отвлечен мост .
37355	Jack nerozumel po španielsky , ale v okamihu vycítil , že jeho obhajca sa usiluje o úspešný priebeh , a spokojne sa usmial . Hľa , čo to značí mať dobrého advokáta ! Don Julian Martínéz y Carvajal sa majestátne zahallil do talára .	Джек , който не разбираше испански , но схвана веднага , че защитникът му правеше сполучлив ход , се усмихна довольно . Ето какво значи <b>добър</b> адвокат ! ... Дон Жулиан Мартинес и Карвахал се загърна величествено в тогата си .
43276	- Pracuje na objave vakcíny proti škvrnitému týfu a nemohol by vám byť prospešný . - Tak teda , - povedala Fanny hrdinsky , uplatňujúc britskú zásadu organizovaného ústupu , - nepodujali by ste sa na moju prípravu vy ? Odpustte mi , oče , ak som príliš bezočivá , že vyslovujem takúto žiadosť .	Работи върху ваксината срещу петнистия тиф и не ще може да ви бъде полезен . — Тогава — каза Фани героично , прилагайки първото правило за отстъпление в <b>добър</b> ред , — тогава не бихте ли могли да ме подготвите вие ? ... Простете , оче !
54559	- povedal Muriet , keď mních odišiel . - Ale to je dobré znamenie ! - Pračo ?	— Бяга от тебе като от дявола ! — каза Мюрие след излизането му . — Но това е <b>добър</b> признак ! — Защо ? — тъжно попита Фани .

Fig. 3. Concordances of the Bulgarian adjectives *добър* in the corpus

In addition, the corpus is useful as a language material for bilingual lexical and terminological databases and on-line dictionaries development [DK09, ŠGD09]. The Bulgarian–Slovak corpus could be also useful to linguists-researchers for research purposes, e.g. in comparative and contrastive studies of Bulgarian and Slovak (Fig. 3, 4).

	Заявка	
101645	... , blízko sejšba vsadené oči horeli živočišnou zvedavosťou . Žena sa chcela s Fanny pohovárať , ale keď narazila na nepriateľský pohľad , odišla smerom k námestiu , kde dúfala , že nájde lepšie miesto na pozeranie . Zástup jednotaj rástol a bol čoraz vzrušenejší .	Черните ѝ тясно поставени очи горяха от животинско любопитство . Жената беше готова да разговаря , но като срещна враждебния поглед на Фани , отмина към площада , където се надяваше да намери по - хубаво място за гледане . Тълпата се увеличаваше и ставаше все по - възбудена .
101911	Fanny zrazu pochopila , že z námestia už nebude môcť odísť . Zástup veľmi zhustol a nikto nechcel opustiť miesto , ktoré si obsadil . Vyhrážne pokusy falangistov v uniformách preraziť si cestu k námestiu ostali neúspešné .	Изведнъж Фани разбра , че не можеше да се отдалечи от площада . Тълпата беше станала много гъста и никой не искаше да напусне заетото вече място . Опитите и заплашванията на една група фалангисти в униформа да си пробият път към площада останаха безуспешни .
101953	Výhrážne pokusy falangistov v uniformách preraziť si cestu k námestiu ostali neúspešné . A aj ulica , po ktorej mali priviezť odsúdených , bola taká upchatá , že vojaci napriek nadávkam nevedeli dav rozohnať , a tak ho nechali rásť . Fanny došla k pohyblivému lešeniu s drôtenou sieťou , ktorým mohli v prípade potreby ulicu uzavrieť , a tak sa ocitla pri robotníkoch , čo s trpkosťou pozorovali zástup .	Опитите и заплашванията на една група фалангисти в униформа да си пробият път към площада останаха безуспешни . Дори улицата , по която щяха да докарат осъдените , беше тъй задръстена , щото войниците въпреки ругатните си не сполучиха да разпръснат множеството и го оставиха да приижда . Фани се приближи до една подвижна скеля с телени мрежи , приготвена да затвори в случай на нужда улицата , и така се озова при работниците , които горчиво наблюдаваха тълпата .
102216	Odkiaľsi prišla rota dobrovoľníkov , aragónskych dedičanov , a načisto upchala námestie . Panebože , aké to bolo všetko odporné ! Od zástupu razilo potom a olivovým olejom .	Отнякъде пристигна цела рота доброволци – арагонски селяни – и задръсти напълно площада . Господи , колко отвратително беше всичко ! ... От тълпата се разнасяше миризма на пот и зехтин .
102815	Ľudia žiadostiví divadla natahovali krky a dvíhali sa na prsty . Každý chcel vidieť odsúdených , lepšie im pozrieť do tváří a vychutnať živočišnú radosť , že nie je na ich mieste . Dav ovládlo všeobecné vzrušenie .	Жадните за зрелище хора протегнаха шияте си и почнаха да се повдигат на пръсти . Всеки искаше да види осъдените , да се взре по - дълбоко в лицата им и да изпита животинската радост , че не беше на тяхно място . Една всеобща възбуда завладе тълпата .
102957	Medzitým sa nákladniak s odsúdenými pomaly blížil k popravisku . Dav dosiaľ ovládala len veľká zvedavosť , ktorá ho ochromovala a nútila mlčať , ale zrazu si spomenul , že musí voči odsúdencom prejavovať nenávisť . - Smrť komunistom !	Между това камионът с осъдените приближаваше бавно мястото на екзекуцията . Досега тълпата беше обладана само от парливо любопитство , което я поразяваше и караше да млчи , но изведнъж си спомни омразата , която трябваше да изпитва към тях . - Смърт на комунистите ! ... – извика един фалангист с превързана ръка .

Fig. 4. Concordances of Bulgarian verb *беше* – (imperfect form of verbs of imperfective aspect in Bulgarian) of the verb *съм* (Dimitar Dimov's *Doomed Souls* and the Slovak translation thereof)

Further work will involve enriching the annotation of the corpora that will increase their representativeness, with regard to its composition and size. We will evaluate existing possibilities for automatic morphosyntactic annotation and lemmatisation of Bulgaria texts and for automatic syntactical annotation of both Slovak and Bulgarian texts of the corpus.

## 7 Conclusion

The paper describes Bulgarian-Slovak and Slovak-Bulgarian parallel and aligned corpora. The corpus is aimed as a widely-distributable dataset for language engineering, for the development of machine translation and other multilingual technologies, such as tools for the development of lexical databases and digital dictionaries, for human translation training and contrastive and terminology studies. Although intellectual property right prohibit open distribution of the corpus, we try to make it as accessible to the wide scientific community as possible, either by providing statistics or language models on request.

## References

- [DEI<sup>+</sup>98] L. Dimitrova, T. Erjavec, N. Ide, H. Kaalep, V. Petkevič, and D. Tufiş. Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In *Proceedings of the COLING-ACL'98*, pages 315–319, Montréal, Québec, Canada, 1998.
- [DK09] L. Dimitrova and V. Koseska–Toszewa. Bulgarian-Polish Corpus. *Cognitive Studies/Études Cognitives*, 9:133–141, 2009.
- [DKRR10] L. Dimitrova, V. Koseska–Toszewa, R. Roszko, and D. Roszko. Application of Multilingual Corpus in Contrastive Studies (on the example of the Bulgarian-Polish-Lithuanian Parallel Corpus. *Cognitive Studies/Études Cognitives*, 10:217–240, 2010.
- [IBR00] N. Ide, P. Bonhomme, and L. Romary. XCES: An XMLbased Encoding Standard for Linguistic Corpora. In *Proceedings of the Second International Language Resources and Evaluation Conference.*, pages 825–830, Paris, 2000. ELRA.
- [R. 06] R. Garabík and В. П. Захаров. Параллельный русско-словацкий корпус. In *Труды международной конференции Корпусная лингвистика*, pages 81–87, Санкт-Петербург, 2006. Издательство С.-Петербургского университета.
- [Sch97] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In Daniel Jones and Harold Somers, editors, *New Methods in Language Processing*, Studies in Computational Linguistics, pages 154–164. UCL Press, London, GB, 1997.
- [ŠGD09] M. Šimková, R. Garabík, and L. Dimitrova. Design of a multilingual terminology database prototype. In V. Koseska, L. Dimitrova, and R. Roszko, editors, *Representing Semantics in Digital Lexicography. Proceedings of the MONDILEX Fourth Open Workshop*, pages 123–127, Warsaw, Poland, 2009.
- [VG07] D. Vasilišínová and R. Garabík. Parallel French-Slovak Corpus. In *Computer Treatment of Slavic and East European Languages. Proceedings of the conference Slovko 2007*. Tribun, Brno, 2007.
- [VNH<sup>+</sup>05] D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, and V. Nagy. Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing*, pages 590–596, 2005.

# The Slovak Dictionary of Collocations

Peter Ďurčo

Faculty of Arts, University of St. Cyril and Methodius, Trnava  
Pedagogical Faculty, Comenius University, Bratislava

**Abstract.** The principles of the first Slovak dictionary of collocations are described here. The dictionary is based on the data of the Slovak National Corpus and it contains 250 collocational profiles of the most frequent Slovak nouns.

The Dictionary of Collocations is the first of its kind in Slovakia. Some countries have already published such types of dictionaries a couple of years ago. They are issued periodically, and their titles vary depending on the countries' respective lexicographic tradition. These are, for example, dictionaries of collocations in English, in Russian dictionary of word's collocability (Словарь сочетаемости слов), style guides in Polish (Słownik Dobrego Stylu) or dictionaries of words and phrases (Wörter und Wendungen, Stilwörterbuch) in German. However, all these dictionaries are hybrid ones when it comes to their type. Apart from collocations, they comprise, to a varying extent, the elements contained in explanatory dictionaries: multiple meanings listed under a key word, definition of single words' meanings lacking collocational ones, inclusion of grammar rules and stylistic qualifiers of base word. Moreover, both a classification system applied to collocations and the examples of collocations given under separate parts of an entry are not always consistent and clearcut. Likewise, there have not been precisely pre-determined criteria governing the selection of collocations to be included into respective dictionaries that frequently give just a fragmentary picture of word collocability.

A special type of Slovak collocational dictionary has not yet been compiled. Collocations are naturally listed in explanatory dictionaries serving the dual purpose of: exemplifying the meaning of a word and providing the examples of multi-word nominations, such as multiword terms, lexicalized multiword units, phraseological units and multi-word proper names. However, the denominative function of collocations is not the only application of the collocability of a word. The majority of words have their standard, usual or typical collocability as a main feature. Each word in lexicon has its own systemic (and, of course in communication, its updating) potential to enter into meaningful combinations with other words; it has its own unique collocability (extensionally defined set of commonly co-occurring elements, that may meaningfully co-occur) and compatibility (intensionally conditioned paradigmatic classes of elements, that may meaningfully co-occur with one another), which define a standard, lexical distribution of a word, and create the so-called collocational paradigm (a collocational profile) of a word – an unique feature of each word in lexicon. The affinity of collocational profiles is a proof of the semantic affinity of words within the framework of their lexical and semantic paradigms, and shows its extent. The comparison of collocational profiles of two words separately and when they co-

occur with one another also gives evidence as to whether it is a free or fixed word combination.

Hold generally, each word has a certain collocational potential, ranging from monocollocability (i.e., collocability with a single element, which is usually easily identifiable, e. g. *úhlavný nepriateľ*, *ceriť zuby*, *mražiť čelo*, *dvere dokorán*), through fixed semantic collocability (e. g. *domová* | *drevená* | *dvojitá* | *dvojkřídlová* | *futbalová* | *garážová* | *vchodová* | *vstupná* | *zamknutá brána*), or through the fixed lexical or idiomatic collocability (e. g. *brána do neba* | *brána do pekla* | *brána do raja* | *brána do srdca* | *brána do sveta* | *brána do života*, *hrať na bránu* | *ohrozovať bránu* | *opečiatkovať bránu* ...), that may be identified, in particular, on the basis of its denominative function, or due to a special meaning attached to a multiword expression as a whole, that simultaneously establishes stability of the collocation itself, up to the so-called typical or usual collocability that is materialised through using a word in texts, and it is identifiable by means of the frequency distribution of elements co-occurring with the given word (e. g. *absurdný* | *akčný* | *alegorický* | *anekdotický* | *animovaný* | *autentický* | *autobiografický* | *baladický* | *banálny* | *biblický príbeh* | *príbeh podľa skutočnej udalosti* | *príbeh bez konca* | *príbeh lásky* | *príbeh so šťastným koncom*; *široká* | *školská* | *tážká* | *úzka* | *velká* | *vysoká* | *zadná* | *záhradná* | *zámocká* | *železná ... brána*).

Various stochastic methods are used in corpus linguistics to ascertain the distribution of the elements co-occurring with a word, so as to eliminate imperfection caused by relying on a simple frequency distribution of elements. However, various statistical methods for eliciting probability distribution of two co-occurring elements within a certain group also frequently give a disparate image, and they provide the sets that differ from each other both with respect to the list of their elements and to their significance and ranking on the list.

Both the extend of extensional set and the character of intensional class of elements in the neighbourhood of examined word are of arbitrary nature, and although not always semantically motivated, they can only be conditional upon the preferential usage in the collocability of a lexeme, upon the reference relationship between two lexemes, or just upon a high common text frequency. Moreover, there is an anomalous or irregular combinatorics of a word caused by the dropping out of a lexical element from its standard paradigmatic and syntagmatic relations. If this is the case, it is an idiomatic collocability of a word that is functionally restricted to a couple of elements, or just to a single isolated element, e.g., necrotisms that evince the highest possible level of disintegration of lexical and semantic paradigms and also, as a rule, of word form paradigm. In collocational paradigms, words do not enter into direct, easy to envisage or even predictable relations, in which neither analogy nor even reversibility of relations works.

Through this frequency stability, the high frequency of a certain co-occurrence of words may exercise an influence upon the stabilization of a meaning attached to an expression as a whole. Thus, we may perceive the collocability of a word as the category that heads forward from the centre comprising the elements having a high degree of collocability towards the periphery consisting of isolated elements having no potential of combination. However, the centre and the periphery may be perceived otherwise, namely that the former (the centre) comprises the elements with obviously

restricted collocability, and the latter (the periphery) is composed of the words of very wide, fuzzy and not exactly restricted collocability.

Although the co-occurrence frequency of two elements has definitely affected the perception of its invariability, it does not lead to the conclusion that a fixed phenomenon automatically means a frequent one. Phraseology has proven so, as the limited or individual collocability of particular elements is their underlying quality, while, in terms of their frequency, these units may stand in the periphery of the word-stock. Word's collocability must be therefore understood, and also examined and recorded as a category that has the nature of a continuum, from fixed expressions through common / typical / frequent ones up to the so-called 'unfixed' expressions that are mostly used as topical word phrases occurring in a new and unexpected context.

When examining collocations, scepticism arises from two essential issues. Firstly, it is not reliably possible to subdivide the whole continuum into free, typical, common, conventional, lexicalized and, in varying degrees, phraseological or idiomatic collocations. Secondly, the number of significant word co-occurrences increases depending on the absolute occurrence frequency of a word / word form, and it is not possible to set the threshold or limit of frequency distribution for a base word in terms of free and fixed combinatorics of its collocates.

The complexity of examining the collocability also results from the fact that the set of collocations consists of diverse items ranging from free, highly-frequent co-occurrences of two, three or more words with no apparent connection to each other or with restricted collocational range (e.g. *vstúpiť do miestnosti / do svedomia niekomu*, but *stúpiť do blata* or *nastúpiť do autobusu*), up to the collocations that, although not statistically significant in terms of their frequency distribution, represent one structural or semantic unit (e. g. *jemne / laický / ľudovo / obrazne / slušne / úprimne / zjednodušene / ... povedané*). Although the frequency and stability are basically related to each other, it is necessary, instead of relying on the 'ranking' of collocates at a certain scale, to choose, as a starting point, the model of a certain gradual radial structure of words and their collocates that represents a fuzzy and vague set of items, where the perception of the centre and periphery may vary depending on whether we take into account a quantitative criterion (co-occurrence, or frequency distribution of elements) or a qualitative one (meaning attached to a word combination as a whole).

Paradoxically though it may seem, both the combinations of words with restricted collocational range that rarely stand isolated, and the highly frequented co-occurrences of highly frequented words may be deemed to be fixed collocations. Between these two extremes, there exists virtually an undividable set of the most varied distribution relationships exhibiting lower or greater frequency differentiation. That is why all the statistical measures, based essentially on the assumptions of the expected and actual co-occurrence of elements in a set, always give different results and identify diffuse subsets of elements having different preferences. On the other hand, the identification of word expressions in the text is made harder by the fact that the elements of fixed combination do not necessarily stand next to each other, but at different distances, often exhibiting a different word order patterns. This applies especially to phraseological combinations that, in addition, have multiple variants.

From a psychological point of view, the collocations are word associations that are spontaneously acquired by a speaker during language acquisition. Associative

links that are being formed become a natural part of the speaker's lexicon in his semantic memory, and they are retrieved from the memory as ready-made units. This has also been confirmed by psycholinguistic experiments verifying spontaneous reactions and measuring the response times of associations elicited by verbal stimuli.

In terms of language variation, associations have not clearly-defined boundaries, but their networks overlap with each other forming a continuum ranging from a variety of quantitative (frequency) and qualitative (semantic) preferences to unambiguous semantic differences. Although we can in Slovak *dať návrh / otázku / odpoveď / žiadosť*, and although the word *dať* (to give) has the meanings 1. poskytnúť (to provide), 2. predložiť (to present), 3. položiť (to lay), out of these semantic equivalents we can *poskytnúť* only *odpoveď* (an answer), *Návrh* (a suggestion) and *žiadosť* (a request), we can only *predložiť* (to present) and in the meaning of "to give" the verb *položiť* (to lay) refers only to the noun *otázka* (a question). On the other hand, these seemingly cognate verb bases develop their own differentiated collocation paradigms displaying a unidirectional relationship to their hyperonym and having basically nothing in common with other word bases when it comes to the most frequently occurring collocational relations.

What is essential about collocability is the fact that it is not linked to a lexeme, but to its specific meaning. A word has dichotomous relationships depending on its occurrence within the system or within the context what, in an over-simplified way, means that since the meaning of a word is context-based, its collocability constitutes its meaning, but the text simultaneously proves the context dependence of meaning.

What is further essential about collocability of a word is the fact that it is dependent on word morphology, i.e. collocates of the word are not primarily associated with a lexeme as a whole. Collocability is linked, in varying degrees, to different forms of the word paradigm. Each lexeme has a unique combinatorics of its word forms with the collocates.

Depending on the planned dictionary functions, the collocations may be described in different ways. With regard to collocation dictionaries, there is no one ideal solution.

Our dictionary is based on the premise that a basic structure is the two-word combination of a base word form and a collocate. Another premise implies that the word collocability is a fuzzy set with a quantitatively and qualitatively definable centre, and unlimited periphery out of the fixed, typical and innovative word combinatorics. The centre is defined primarily as a category of phrases having denominative function; the significance of frequency distribution of word combinations / word forms determines their different position between the centre and the periphery.

The Dictionary of Collocations has not been intended just to list and describe multi-word lexicalized and phraseological units, but most importantly to list the so-called typical or conventional frequently occurring co-occurrences, which means also the words with relatively wide collocability that is differentiated in terms of frequency, and thus it may be classified. The dictionary entries include irregular system collocations<sup>1</sup> (idioms and phrasemes), common regular text collocations

<sup>1</sup> See the classification of word combinations in Čermák, František: *Kolokace v lingvistice*. In: František Čermák – Michal Šulc (eds.): *Kolokace. Studie z korpusové lingvistiky*. Sv. 2. Praha, Nakladatelství Lidové noviny, Ústav Českého národního korpusu 2006, 9-16.

(*zimná rekreácia, letné prázdniny*) and conventional text-system collocations (*krájať nadrobno, hovoriť úsečne, vystúpiť z auta*). Regular system collocations of terminological (*difúzna množina, pravý uhol*) and proper type (*Vysoké Tatry*) have not been included. When choosing this approach, in terms of lexicography, the importance is attached to a methodological starting point, namely that collocations are not primarily linked to the abstract notion of a lexeme, respectively a lemma, but they are linked to word forms. In our Dictionary, the collocations are arranged by the word form of a keyword – key noun in combination with other noun, verb, adjective or pronoun. Within separate structural types of collocations are the collocations listed alphabetically.

The Slovak National Corpus of the Ľudovít Štúr Institute of Linguistics of the Slovak Academy of Sciences was used to conduct the research (the first dictionary compilation phase – approx. 350 million text words, the second one – approx. 530 million text words). To detect collocation profiles, it was used Sketch Engine, a software tool for the analysis of word combinations forms. Existing dictionary sources were permanently consulted during the entire process of collocation profiles creation. These include also less frequent, but fixed lexicalised and phraseological expressions that may escape detection by statistical analysis of the Corpus.

The main entries matched a selection criterion of the absolute frequency of words in the Slovak National Corpus. There have been compiled 250 most frequent Slovak nouns that simultaneously represent a certain lexical minimum for those interested in learning Slovak as a foreign language. The biggest collocational profiles have the following words: *svet* (world; 599 collocates), *charakter* (492), *oko* (eye; 488), *práca* (labour, work; 427), *problém* (422), *miesto* (place; 408), *rok* (year; 407), *cesta* (way; 403), *strana* (side, party; 402), *život* (life; 456). The smallest collocational profiles have the following words: *okres* (district; 87 collocates), *domácnosť* (household; 86), *cudzina* (foreign country; 84), *istota* (certainty, safety; 82), *parkovisko* (parking place; 82), *jeseň* (autumn 78), *konzumácia* (consume; 77), *zima* (winter; 77), *spisovateľ* (writer; 71), *jar* (spring; 64), *narodeniny* (birthday; 64), *letisko* (airport; 62), *informatika* (54), *kára* (cart; 40).

A target group of this dictionary includes all those who actively work with the language, and seek associative hints for their written texts or spoken utterances. They may be writers, translators, journalists, editors, teachers, students and even linguists themselves, especially lexicographers.

# Generating Sets of Synonyms between Languages

Ondrej Dzurjov<sup>1</sup>, Ján Genčí<sup>1</sup>  
and Radovan Garabík<sup>2</sup>

<sup>1</sup>Department of Computers and Informatics, Technical University of Košice  
<sup>2</sup>Ludovít Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava

**Abstract.** Princeton WordNet is a lexical database that contain sets of synonyms for the English language together with their semantic relationship. In this paper we explore several methods of generating synsets in another language by using English language WordNet and a bilingual dictionary. The methods have been used to generate Slovak language synsets and to bootstrap a Slovak WordNet database.

## 1 Introduction

Communication is one of the most important things in the social life of every person. It is important to understand other people, learn or change experience. Nowadays, people travel all over the world and learn from books written in foreign languages, so the presentation of knowledge in more languages becomes a common fact. But learning new words is not enough to translate a sentence properly. The key is to know the meanings of a word, how to use it and where to use it.

There are many projects that offer multilingual synonym dictionaries with semantics included like BalkaNet [1], EuroWordNet [2] and Global WordNet [3]. All these projects have something in common. They are based on English Princeton WordNet [4]. English WordNet is a lexical database that connects sets of English synonyms into a semantic net. Chapter one of the paper deals with it in detail. Every project mentioned above connects similar meanings for several languages, but not for Slovak. This paper describes some methods on how to automatically generate Slovak synonyms using the WordNet database and online Slovak-English dictionaries. Generating groups of Slovak synonyms from English ones ensures that Slovak synonym sets will be properly connected to the English equivalents. The article contains statistics records about the quantity of results and an experiment that shows how a quantity can be influenced by the usability of words in an English synset.

Finally, after generating, a small Slovak WordNet was created by using generated Slovak synsets and English WordNet as a pivot. This paper also describes an approach to the building of Slovak WordNet.

The presented work is based on two previous projects [5] and [6]. The first one presents proof of concepts for the automation of Slovak synset generation by using online dictionaries and the second one focuses on building equivalent bilingual synsets from dictionary items only.

## 2 Introduction of WordNet

WordNet is a lexical database of English synonyms containing nouns, adjectives, verbs and adverbs. The development began in 1990 at Princeton University. WordNet has two main characteristics:

- Words with the same meaning are grouped into synsets – sets of synonyms.
- Synsets are connected by relations and create a synonymic net of synsets.

Today, the WordNet database is at version 3.0 and contains more than 117 000 synsets. Table 1 shows some statistical information about the WordNet database.

Part of speech	Nouns	Adjectives	Verbs	Adverbs	Totals
Unique strings	117798	21497	11529	4481	155287
Strings with one sence	101863	16503	6277	3748	128391
Word-meaning pairs	146312	30002	25047	5580	206941
Synsets	82115	18156	13767	3621	117659
Synsets of one word	42054	11353	8041	2400	63848

Table 1. WordNet 3.0 statistics

The basic relations between synsets in WordNet are:

**Synonymy** – is a relation between literals (synonyms) in one synset.

**Hyponymy** – is a relation of sense specification between synsets, a relation heads from general synset to a more specific synset (motor vehicle → car, automobile).

**Hypernymy** – is a relation of sense generalization between synsets, a relation heads from a specific synset to a more general synset (motor vehicle ← car, automobile).

**Meronymy** – is a relation between a term denoting the part and a term denoting the whole, leading from the whole to its part (car, automobile → engine).

**Holonymy** – is a relation between a term denoting the part and a term denoting the whole, (car, automobile ← engine).

## 3 Generating Slovak synsets

### 3.1 WordNet and Slovak-English dictionary

The process of building Slovak synonym sets uses the WordNet database as a source of English synsets. The most important WordNet relations used for generating Slovak synonym sets are synonymy, hypernymy and holonymy, where synonymy is the equivalence in meaning (for words in the same synonym set). Hypernymy and holonymy are described in section 2. Generated Slovak synsets are mapped to their English equivalents, so after the process of building Slovak synsets, WordNet relations should be valid also between Slovak synsets.

The second very important source of data is a good quality electronic English-Slovak dictionary, for example an online dictionary which is used for search of Slovak synsets. The size of a translator's database and quality of its translations are very important for the quantity and quality of created groups of Slovak synonyms.

### 3.2 Methods for generating Slovak synonyms

#### Method A

This method uses a synonym relation between words in one synset. When we translate English synonyms, we can expect that some translations will contain the same words. So the words that are in two or more translations constitute a synset in the Slovak language.

For example:

We have the English synset *{kind; sort; form; variety}*, which means “a category of things distinguished by some common characteristic or quality”.

After translating using English-Slovak dictionary [7], we get these groups of Slovak words:

*kind* – druh, rod, kategória

*sort* – druh, akosť, trieda, typ, forma, chlap

*form* – forma, tvar, podoba, formulár, blanketa, formula

*variety* – rozmanitosť, odroda, výber, druh, rad, množstvo, mnohotvárnosť, rôznosť

After intersecting all the pairs of translations, the final group *{druh, forma}* represents the Slovak equivalent of the English synset *{kind; sort; form; variety}*.

Advantages of this method: sense accuracy

Disadvantages of this method: empty Slovak synsets for English synsets consisting of one word

#### Method B

The next method is based on an idea that English words with one sense should be translated into one group of synonyms. If a synset contains more words with one sense, these words should have similar translations. A Slovak synset will be created by the union of these translations.

The synset *{kind; sort; form; variety}* in the previous example contains one univocal word:

*kind* – 1 sense, translation: druh, rod, kategória

*sort* – 4 senses

*form* – 16 senses

*variety* – 6 senses

After the translation of all univocal words and the union off all these translations we get the Slovak synset: *{druh, rod, kategória}*.

Advantages of this method: possibility of creating Slovak synsets from English synsets consisting of one word (this word must have only one sense in WordNet)

Disadvantages of this method: quality of Slovak synsets depends on translation accuracy, univocal words in English can have more senses in Slovak; empty synsets for English synsets with no univocal word

### Method C

Method C alsoises hypernym and hyponym synsets in addition to the default English synset. There are small differences in sense between some English synsets that are in a hypernymic or hyponymic relationship. It is expected that after translation some words will be the same for more synsets. In this method two groups of words are created. A group of words belonging to the default synset and a group created from all its hypernyms and hyponyms. The next step is to translate these groups and then to intersect them.

For example, we have a synset *{kind; sort; form; variety}* (group 1). Its hypernyms and hyponyms will create one group: *{category, type, brand, genus, species}* (group 2).

Translated groups:

Group 1 – druh, rod, kategória, akosť, trieda, typ, forma, chlap, tvar, podoba, formulár, blanketa, formula, rozmanitosť, odroda, výber, rad, množstvo, mnohotvárnosť, rôznosť

Group 2 – kategória, skupina, trieda, typ, symbol, litera, druh, odroda, značka, označenie, známka, kvalita, akosť, ohorok, rod, forma, tvar

The final Slovak synset will be: *{druh; rod; kategória; akosť; trieda; typ; forma; tvar; odroda}*

Advantages of this method: possibility of creating Slovak synsets from English synsets consisting of one word; quantity

Disadvantages of this method: lower quality of Slovak synsets

### Method D

This method is a modification and extension of method C. It is also based on small differences in sense between WordNet synsets. This method doesn't use a synset that we are using for generating its Slovak equivalent. The aim is to create an intersection between the translation of its hypernym synset and the translation of its holonym synsets, so ultimately the synset equal to the default English synset should be created. At first, two groups are created: the first group represents the hypernym synset (more general), the second group represents the union of all hyponym synsets (more specific). The Slovak synsets are created by translating and intersecting these groups.

For the synset *{kind; sort; form; variety}*:

Hypernym group: *{category}*

Hyponym group: *{type, brand, genus, species}*

Translated groups are:

Hypernym group: kategória, skupina, trieda

Hyponym group: typ, symbol, litera, druh, odroda, značka, označenie, známka, kvalita, akosť, ohorok, druh, rod, skupina, trieda, forma, tvar

The final Slovak synset is: {skupina, trieda}

Advantages of this method: possibility of creating Slovak synsets from English synsets consisting of one word

Disadvantages of this method: small quantity; lower quality of Slovak synsets

## 4 Statistics of results

All previous described methods were used for generating Slovak synonym sets. The whole process was divided into four steps:

1. translation of English words from WordNet
2. using methods A-D to build Slovak synsets from translations according to English words in English synsets
3. additional correction of created synsets (removing words with duplicate entries and words with incorrect parts of speech)
4. storing a new synset with reference to an English equivalent

After the process of generation, statistics of the results were created to evaluate the reliability of automatic generation for all presented methods.

### 4.1 Complete Results

The next table shows results for an attempt to generate Slovak synsets for a complete WordNet database.

	<b>Totals</b>	<b>Nouns</b>	<b>Adjectives</b>	<b>Verbs</b>	<b>Adverbs</b>
<b>Synsets in WN</b>	117659	82115	18156	13767	3621
<b>Total EN synsets with Slovak synset</b>	40521 (34.4%)	26787 (32.6%)	6859 (37.8%)	5839 (42.4%)	1036 (28.6%)
<b>Method A</b>	10267 (8.7%)	5705	2175	2109	278
<b>Method B</b>	30243 (25.7%)	20510	6059	2715	959
<b>Method C</b>	11533 (12%)	8192	-	3341	-
<b>Method D</b>	1917 (1.4%)	1348	-	569	-

**Table 2.** Comparison of synset generation methods

Slovak synsets were generated for 34% of all English synonym sets in WordNet. Method A generated less than 9%. It is because most of the synsets in WordNet contain only one word. A high amount of univocal words in WordNet caused the generation of more than 25% of synsets with method B. Method D generated much fewer synsets than other methods so it is not as effective as expected.

Adjective and adverb synsets have not a hypernymy and hyponymy relationship between them so methods C and D could not be used in this case.

## 4.2 Experiment

Low quantity of generation was caused by more factors:

- many words could not be translated because there was no translation in the dictionary for them
- most English synsets consist of one word
- absence of some relations used for generating

We created an experiment to get some statistical information for commonly used words. We created a group of 300 words randomly selected from the 5000 most used words in English [8]. Then we found English synsets containing these words (1709 synsets). The next table shows data from this sample of commonly used words.

	<b>Totals</b>	<b>Nouns</b>	<b>Adjectives</b>	<b>Verbs</b>	<b>Adverbs</b>
<b>Synsets in a sample</b>	1709	769	397	429	114
<b>Total EN synsets with a Slovak synset</b>	946 (55.4%)	491 (63.9%)	171 (43.1%)	237 (55.2%)	47 (42.2%)
<b>Method A</b>	559 (32.7%)	255	126	145	33
<b>Method B</b>	404 (23.6%)	201	97	67	39
<b>Method C</b>	505 (29.6%)	337	-	168	-
<b>Method D</b>	112 (6.6%)	67	-	45	-

**Table 3.** Statistics of sample synsets generation

There are some important numbers in this table in comparison to the results in table 2. Generating common synsets is much more successful. 32% of generated synsets are by method A which is the biggest increase out of all methods. Method B is almost at the same value which could be caused by the balanced location of univocal words in WordNet synsets.

## 5 Building the Slovak WordNet

### 5.1 Automatic synset building

The approach described above has been used to bootstrap a basic Slovak-English-German-Polish-Lithuanian dictionary<sup>1</sup>. The Slovak-English synset pairs have been generated as described before, the other languages have been pre-filled from other sources and then manually proofread. The Slovak part of the structure then serves as a base for a small Slovak language WordNet.

We selected the ten thousand most frequent words from the Slovak National Corpus (balanced subcorpus prim-4.0-vyv). We then generated synsets for each of the noun, verb, adjective and adverb categories of these words. A web-based application is used to further edit the generated synsets and their relation to the English synsets. The application allows for general M:N mapping between English and other synsets – the English WordNet serves as a pivot language in the dictionary, even if the external appearance will be that of a *Slovak*→*other language* one. An additional link can be specified between synsets in other languages and Slovak synsets inside a set of synsets linked to the same English synset. This is used in job titles or animal nomenclature, where the (usually) gender-neutral English noun has two Slovak synsets assigned, one masculine<sup>2</sup> and one feminine. German, Polish and Lithuanian nouns (which mostly keep the same distinction as Slovak) are then linked with the corresponding Slovak synset.

### 5.2 Synset structure

Each synset has an optional gloss in its own language (parallels the English WordNet structure) – used only if further explanation or refining of the sense is desired.

There are several possible marks applied in the (non English) synset description:

- One or several constituent words in the synset can be marked as “major”, giving it a distinct visual realization in the final dictionary version.
- The whole synset can be marked as “imprecise”. This is used in cases where there is no direct semantic equivalent to the English synset, but the synset had to be filled in, most likely because it was a hypernym of other existing synset(s). This is mostly present in concepts that are realized in other languages as phrases or descriptions (e.g. the English noun *uxoriousness* has no Slovak language equivalent as a noun describing the trait – the meaning combines two rather different concepts, the verbal construction *byť pod papučou* and a dative noun phrase *oddanosť manželke*)
- Individual words in the synset can be marked as “unsure”. This is purely a temporary measure for the editor to record that he or she was unsure about the equivalence or meaning and the synset has to be re-checked later.

<sup>1</sup> Sponsored by the Slovak Online (Lifelong Learning Programme 504873-2009-LLP-SK-KA2-KA2MP) project.

<sup>2</sup> Strictly speaking, a Slovak masculine noun should be assigned into two different synsets, a general one encompassing both genders (or gender agnostic) and a strictly masculine hyponym. However, we considered this distinction too detailed for the purpose of the database.

Additionally, a synset in the database can be marked as “checked” (by an independent reviewer).

### 5.3 Verbs

Links between other parts of speech are straightforward; there are only a few isolated cases where the situation is more complicated (such as the inclusion of numerals as nouns, or English adverbs whose Slovak equivalents are classified as particles). On the other hand, verbs are more complicated. Features that deserve special care are negation, aspect and reflexivity.

Verb negation in Slovak is accomplished (with very few exceptions) by prefixing the verb with *ne-*, which is then seen as a separate, derived verb. We included the most frequent negative lexemes in the database if there was a corresponding English synset (e.g. *disagree*↔*nesúhlasit*); for all other verbs, we have only the affirmative form.

Verb aspect in Slovak is mostly inherent in the lexical level – verbs can be either perfective, imperfective, or ambivalent (which is in fact just the conflation of both aspects into one lexeme), although ways of deriving the opposite aspect exist, such as prefixes turning an imperfective verb into the perfective and morphology root changes to turn a perfective verb into an imperfective one. In the database, we treat perfective/imperfective verb pairs as separate lexemes and assign them to separate synsets that are linked to the same English synset (unless there is a different English synset for the opposite aspect). The presence of both perfective and imperfective verbs inside one synset is prohibited and is automatically enforced by comparing the synsets entered against a list of perfective and imperfective verbs respectively. We do not include Slovak verbs that are only formally derived from the opposite aspect and are not used reasonably frequently in the language. In particular, frequentative/habitual verbs can be derived almost mechanically, but only the frequently used ones are included in the database.

Verb reflexivity is realized with special reflexive pronouns *sa*, *si* that are considered part of the lexeme, although they are written separately from the verb proper and their position in the sentence varies and can be quite remote from the verb itself. If there is a Slovak reflexive/non-reflexive verb pair and the meaning of both of the verbs corresponds to one English synset, both the reflexive and non-reflexive verbs are assigned to two different synsets linked to the same English synset, often separately for perfective and imperfective aspects (therefore producing in some cases four different Slovak synsets linked to the same English one).

Part of speech	Nouns	Adjectives	Verbs	Adverbs	Totals
Unique strings	12941	3321	1150	982	18394
Strings with one sense	10239	2305	953	702	14199
Word-sense pairs	18740	5551	1400	1505	27196
Synsets	9317	2329	830	549	13025
Synsets of one word	3916	773	426	141	5256

Table 4. Slovak WordNet statistics (at the time of writing)

## 6 Conclusion

It is clear that it is not possible to generate synsets for the whole WordNet database. Quality and quantity of Slovak synsets correspond to the usability of words in real life. The main problem is that there is no translation in dictionaries for some/many words in the WordNet database.

We used four different methods for generating Slovak synsets and each of them have their advantages and disadvantages, and some Slovak synsets were produced from English ones by more than one method and it is not possible to select the best one automatically. It is also important to balance the quality and quantity of results. For example: Method D uses an idea too complex to find Slovak synsets and the number of results is very low. Also, by joining two or more techniques (described in this article) together we can achieve better results. The output from all methods covered 34.4% of all English synsets.

Generated data were used to bootstrap a Slovak WordNet database. Some generated synsets had an incorrect sense or contained words with the wrong part of speech, but the synsets were manually checked and corrected if needed. A web-based application was created to simplify the whole process of building Slovak WordNet. This application was also used to create a basic Slovak-English-German-Polish-Lithuanian dictionary.

## References

- [1] Stamou S., Kemal O., Pala K., Christoudoulakis D., Cristea D., Tufiş D., Koeva S., Totkov G., Dutoit D., and Grigoriadou M. BALKANET: A multilingual semantic network for the Balkan languages. In Proceedings of the International Wordnet Conference, pp. 12–14, Mysore. 2002.
- [2] EuroWordNet. <http://www.illc.uva.nl/EuroWordNet/>
- [3] The Global WordNet Association. <http://www.globalwordnet.org/>
- [4] WordNet – a lexical database for the English language. <http://www.wordnet.princeton.edu/>
- [5] Lapoš P.: Verification of possibility to build EuroWordNet synsets based on on-line dictionaries. Diploma work. KPI FEI TU Košice. 2005.
- [6] Sudynová M.: Generating tool for dictionary records. Diploma work. KPI FEI TU Košice. 2006.
- [7] Slovník.sk: <http://www.slovník.sk/>
- [8] Word Frequency List: [http://en.wiktionary.org/wiki/Wiktionary:Frequency\\_lists/Project\\_Gutenberg\\_1-10000](http://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/Project_Gutenberg_1-10000)
- [9] Dzurjov O.: Computational linguistics – Generating sets of synonyms between languages. Semestral project. KPI FEI TU Košice. 2010.
- [10] Dzurjov O.: Computational linguistics – Generating sets of synonyms between languages. Diploma project. KPI FEI TU Košice. 2010.

# Czech-Slovak Parallel Corpora for MT between Closely Related Languages\*

Petra Galuščáková and Ondřej Bojar

Charles University in Prague  
Faculty of Mathematics and Physics  
Institute of Formal and Applied Linguistics

**Abstract.** The paper describes suitable sources for creating Czech-Slovak parallel corpora, including our procedure of creating plain text parallel corpora from various data sources. We attempt to address the pros and cons of various types of data sources, especially when they are used in machine translation. Some results of machine translation from Czech to Slovak based on the acquired corpora are also given.

## 1 Introduction

The Czech language has twice as many users as the Slovak language, resulting in more foreign texts being translated into Czech than Slovak. Czech and Slovak are closely related languages and thus machine translation from Czech to Slovak is a much easier task than translation from a third language to Slovak. If we need to translate some texts from e.g. English to Slovak and these texts are already translated into Czech, it is easier to translate these translations into Slovak.

Depending on the type of machine translation system chosen, large Czech-Slovak parallel corpora may be needed. In any case, such a parallel corpus serves as a good evaluation set.

In the following, various possible sources for acquiring Czech-Slovak parallel corpora are covered. We attempt to describe the pros and cons of each source, especially with respect to the task of training or evaluating MT systems. The subsequent sections are aimed at the usage of the corpora. We describe experiments that we performed with “Moses”, a statistical machine translation system that was trained and tuned with the acquired corpora.

## 2 Tools

Our ultimate goal was to acquire plain text aligned Czech-Slovak sentences. Therefore, the data that we collected required processing. The first step was segmentation into sentences. We used a trainable tokenizer by Ondřej Bojar (Klyueva N., Bojar O. [2]) and adapted it to Slovak for our purposes.

The alignment between Czech and Slovak sentences was found using Hunalign software (Varga D., Németh L., Halácsy P., Kornai A., Trón V., Nagy V. [5]).

High quality segmentation is very important for good alignment. Specifically, it is very important that segmentation works in the same manner for both the Czech and Slovak. For

---

\* The work on this project was supported by the grants EuroMatrixPlus (FP7-ICT-2007-3-231720 of the EU and 7E09003 of the Czech Republic), GAČR P406/10/P259, and MSM 0021620838.

example, if there is a sentence break after an ordinal number and a dot in one language, there should be a corresponding sentence break in the second language as well. Mismatches in sentence segmentation lead to output as illustrated in Figure 1. Instead of 1-1 alignment, Hunalign resorts to 2-1 or 1-2 alignments. While this does not completely disqualify such alignments, these non-matching alignments are (in other cases) often of lower quality and are simply removed in subsequent steps. Thus the mismatch in segmentation together with simplistic subsequent filtering leads to unnecessary data loss.

Alignment Type	Czech Sentence	Slovak Sentence
2-1	— <s> Viktor nevnímal hovor a zmatek ve vagónu.	Viktor nevnímal vravu a zmätok vo vagóne.
2-1	”Pryč ode mne, vy zloto! <s> Co vám udělaly ty kačátka?	„Preč odo mňa, vy lotri! čo vám urobili tie kačičky?
1-2	Stáří 23 let. Zoolingvistka.	Vek dvadsaťtri rokov. <s> Zoolingvistka.
1-2	II/ MODLITBA	II <s> MODLITBA

**Fig. 1.** Several examples of mismatched sentence segmentation leading to non-1-1 alignments. Sentence breaks that required rejoining to achieve the alignment are displayed as “<s>”.

### 3 Sources for Czech-Slovak parallel corpora

We surveyed several sources, of parallel Czech-Slovak data. The sources differed in several ways. Some sources were more useful than others due to the ease of extracting aligned data from them. We sought plain text parallel corpora, that required no manual annotation.

#### – Books

Books in general are a very good source of data, especially for machine translation purposes, thanks to their high quality text and translation. On the other hand, acquiring such data is quite complicated. The Slovak Academy of Science<sup>1</sup> is currently preparing the Czech-Slovak parallel corpus, which is based on books. The use of this corpus is limited due to copyright restrictions. We used an older version of this corpus that contained 118 books in total: 61 Slovak books translated to Czech, 55 Czech books translated to Slovak and two books translated from a third language into both Czech and Slovak. This version of corpora did not contain alignment, therefore we performed the alignment ourselves.

Books are quite difficult to align, because they often consist of long contiguous texts without reference points. Therefore the quality of automatic alignment needs to be controlled. We also found other problems with aligning. Sometimes the translated text was truncated, more often several sentences were compressed into a single sentence, and in several cases whole passages of text were omitted.

<sup>1</sup> <http://korpus.juls.savba.sk>

– **Acquis JRC**

Acquis<sup>2</sup> is a parallel corpus created from texts of European Union, which is freely available. This corpus offers large amounts of parallel data for all pairs of official EU languages including Czech and Slovak. Czech and Slovak texts were created from the translation of a third language, English, in most cases. The main drawback of the Acquis texts is their monotonous nature with large portions of texts often being repeated. This problem is illustrated in Table 1 where the number of all sentences (lines after our sentence segmentation) and the number of unique lines are compared.

Source	Lines Total	Lines Unique	%
Acquis CZ	926082	608086	65.66
Acquis SK	926082	632916	68.34
Books CZ	153478	148705	96.89
Books SK	153478	149152	97.18

**Table 1.** Comparison of the number of all lines (i.e. sentences in our segmentation) to the number of unique lines. Ec-Europa corpus was already deduplicated and therefore is not listed in the table.

Due to the many duplicated sentences, a random subset of the Acquis corpus selected as a test corpus may often contain sentences that are verbatim present also in the remaining “training” data. The results of the (automatic) MT evaluation based on this corpus can thus be overly optimistic if the MT system is trained on this corpus. Another problem is that this corpus is a collection of legislative texts and the vocabulary is somewhat restricted. For these reasons, evaluation based on this corpus cannot be compared to the evaluation based e.g. on books or newspaper articles. Thus, Acquis is a very good source for training data but it should be combined with more disparate sources for the purposes of testing.

– **Ec-Europa-Eu**

Another source that we examined was the website of the European Commission<sup>3</sup>. This website consists of pages in various language mutations, including Czech and Slovak. Sites in various languages differ by the suffix used in their respective URL and pages with the same name should contain the same text. Thus, alignment at the document level is straightforward.

These texts were manually translated, probably from English into other languages. Unfortunately, very often a portion of a page has been left untranslated but nevertheless is presented under the target language label, so Czech and Slovak pages often contain English parts.

We implemented a custom web crawler for downloading these pages. For technical reasons, we downloaded only a subset of all available pages of the site. In total, we downloaded 25737 Czech and 25918 Slovak web pages. The downloaded pages required some initial cleanup work. We removed all HTML tags and corrected the character encoding. Web pages in Czech and Slovak were paired with each other based on

<sup>2</sup> <http://optima.jrc.it/Acquis>

<sup>3</sup> <http://ec.europa.eu/>

their URLs and the parallel ones were segmented into sentences. Duplicated sentences were removed from these web pages afterwards. We decided to run sentence deduplication before alignment because of the amount of (identical) English text inside both Czech and Slovak variants of the page. The remaining sentences were automatically aligned.

#### – **Eur-LEX**

The Official Journal of the European Union<sup>4</sup> may be used as another source. This source offers a huge quantity of data. The data is somewhat similar to the Acquis corpora and so similar problems may be associated with it. Here Czech and Slovak documents were also created as translations from English. Documents in the corpus are in XML format that first required conversion into plain text.

Based on our observation of the document collection, we sorted documents into two types: lists and texts. Sorting was performed automatically based on the average number of words per line. Documents in which the average number of words per line was less than 2.8 were marked as lists. The remaining documents were marked as texts. Next we counted the number of lines in the list documents. If a given document was marked as a list in both Czech and Slovak and it contained the same number of lines in both the Czech and Slovak versions, then this document was marked as a parallel list. This sorting to texts, lists and parallel lists is not very precise, but it proved to be sufficient in most cases. Parallel lists were then aligned line-by-line. Non-parallel lists and texts were aligned by Hunalign software.

The official alignment performed by the publisher of this corpora is expected to be completed in the near future.

#### – **Other possible sources**

Among other sources, we also translated several sentences from WMT<sup>5</sup>. In the future more web pages from the European Union could be used as data sources. Articles from Project Syndicate<sup>6</sup> are sometimes also translated into Slovak, although they are not generally available on the project web page. If there also exists a Czech version of these articles, they could be used as another source. Sometimes, it is also possible to find news from the Czech News Agency translated into Slovak in a Slovak newspaper. This possible source should be further explored. Another possible source could be movie subtitles translated to Czech and to Slovak.

A comparison of the quantities of data acquired from various sources is given in Table 2. Numbers of documents for various source languages are shown in Table 3.

<sup>4</sup> <http://eur-lex.europa.eu/JOIndex.do>

<sup>5</sup> [http://matrix.statmt.org/test\\_sets/list](http://matrix.statmt.org/test_sets/list)

<sup>6</sup> <http://www.project-syndicate.org/>

Source	CZ Words	SK Words	CZ Tokens	SK Tokens	Sentences	Documents
Acquis	20.4 mil	20.6 mil	24.3 mil	24.4 mil	926.1 k	20135
Books	6.6 mil	6.6 mil	8.1 mil	8.1 mil	550.6 k	118
Ec-europa	0.4 mil	0.4 mil	0.4 mil	0.4 mil	24.2 k	1493
Total	27.4 mil	27.6 mil	32.8 mil	32.9 mil	1.5 mil	21746

**Table 2.** Number of acquired words, tokens and sentences from each type of source. The final version of the Eur-LEX corpus has not yet been completed; therefore, we did not include this corpus. We used an older version of corpora created from books than is currently available.

Language	Documents	Sentences
Czech	55	223.6 k
Slovak	61	321.7 k
Other	21630	955.6 k

**Table 3.** The number of documents for various source languages

## 4 Usage of the corpora

The corpora we collected could have wide ranging utilization. We are primarily interested in machine translation from Czech to Slovak. Since we could manage this task, it will be possible to utilize Czech as a pivot language. We could translate English texts manually to Czech and then use an automatic translation system for translation into other languages that are similar to Czech – for example Polish, Russian or Slovak. Additional thoughts on this concept may be found in Hric J., Hajič J., Kuboň V. [1].

Parallel data also facilitate the automatic creation of a Czech-Slovak dictionary. Such a dictionary may find further use in automatic translation systems.

## 5 Czech to Slovak automatic translation

We attempted to use some of the acquired corpora for training and testing automatic translation tool Moses<sup>7</sup> [3]. Initially we exclusively used the Acquis corpus; later we also included data collected from books.

Acquis data were sorted into training, tuning and testing data sets according to the same procedures used by Phillip Koehn in Euro Matrix project<sup>8</sup>. The training set consisted of 926082 sentences and the tuning set consisted of 4107 sentences. We then used books that were manually translated from Czech to Slovak. Alignments were manually checked and only good alignments from the books were used for training and testing. Only 39 books were used in this experiment. A subset of 4000 sentences from the books was randomly chosen as a testing set, another 4000 sentences were randomly chosen as a tuning set and

<sup>7</sup> <http://www.statmt.org/moses>

<sup>8</sup> [http://matrix.statmt.org/test\\_sets/list](http://matrix.statmt.org/test_sets/list)

the rest of the corpus (145478 sentences) was used as the training set. We used all of the books when selecting the tuning and testing sets; therefore, it was possible for the vocabulary that was used in the training set to also appear in the testing set. Due to occasional repetition of sentences in the books, some overlap of the test and training sets was also a possibility. Numbers of sentences from the testing set, that also appear in the training set are listed in Table 4.

Moses was first trained with the Acquis training set and tuned with the Acquis tuning set. Next we used Acquis as a training set and books for tuning. As a third procedure we used books as a training set and Acquis for tuning and as a fourth procedure we used books for both the training and tuning sets. We also tried training on the Acquis corpus merged with books. We used this merged set for training and the Acquis corpus for tuning as a fifth procedure. Finally, we used books for tuning as the last procedure. The test set acquired from books was used for testing in all of the cases. The result of the evaluation can be found in Table 5. A “BLEU” score [4] was used for the automatic evaluation. This metric is based on a comparison of translated segments to a reference translation.

Training Corpus	Identical Sentences	%
Acquis SK	7	0.2
Acquis CS	5	0.1
Books CS	142	3.6
Books SK	122	3.1
Acquis+Books CS	143	3.6
Acquis+Books SK	122	3.1

**Table 4.** Number of sentences in the testing set that also appear in the training set. Corpora were not pre-filtered.

Training/Tuning Corpus	Training Set Sentences	Tuning Set Sentences	BLEU
Acquis/Acquis	708406	3148	0.1808
Acquis/Books	708406	3802	0.2071
Books/Acquis	137027	3148	0.4661
Books/Books	137027	3802	0.4701
Acquis+Books/Acquis	845433	3148	0.4781
Acquis+Books/Books	845433	3802	0.4887

**Table 5.** BLEU Evaluation of Moses for Czech-to-Slovak using various data sources. The test set was the same for all cases and contained 3860 sentences of randomly selected sentences from books. Numbers of sentences were counted after filtering out sentences that contained more than 40 tokens.

The size of the Acquis training set is much larger than the size of the training set created from books. In spite of this, results acquired when we use books for training are much higher. Size of tuning sets for books are similar to the Acquis corpora. Using books exclusively for the tuning also improves the results. The best result is, not surprisingly, achieved with the training set composed of books and the Acquis corpora when we are tuning with books.

The difference between highest and lowest scores is very large. This may be caused by there being a wider range of word forms than are used in books but not as wide a range as in the monotonous EU legislation.

However, we are aware of the fact that we tested with the same books as were used for training (despite the disjoint subset of sentences). The vocabulary in the training and test sets, including e.g. proper names, can be thus unnaturally similar. To obtain a more realistic estimate of MT quality, we plan to test using sentences from new books that are not included in the training data.

## 6 Summary

We described various types of data sources for parallel Czech-Slovak corpora. The initial cleanup of these sources and the necessary steps used to create our parallel corpora were also described.

Additionally, we have given some preliminary results of our machine translation based on the acquired corpora. The results are closely related to the described characteristics of the data sources that we used. We observed a sharp increase in (automatically estimated) MT quality when books were included in the training data. The exact explanation for this has yet to be determined.

## References

- [1] Hric J., Hajič J., Kuboň V. (2000). Machine Translation of Very Close Languages. *Proceedings of the 6th Applied Natural Language Processing Conference*, pages 7–12.
- [2] Klyueva N., Bojar O. (2008). UMC 0.1: Czech-Russian-English Multilingual Corpus. In *Proceedings of International Conference Corpus Linguistics*, pages 188–195.
- [3] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.
- [4] Papineni K., Roukos S., Ward T, Zhy W. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *ACL 2002, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.
- [5] Varga D., Németh L., Halácsy P., Kornai A., Trón V., Nagy V. (2005). Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing*, pages 590–596.

# Slovene Lexical Database

Polona Gantar<sup>1</sup> and Simon Krek<sup>2</sup>

<sup>1</sup> Scientific Research Centre of the Slovenian Academy of Sciences and Arts

<sup>2</sup> Amebis, d.o.o., Kamnik; Jožef Stefan Institute, Slovenia

**Abstract.** The paper describes the concept of the new Slovene lexical database which is compiled within the “Communication in Slovene” project. The database has a twofold goal: it is intended as the basis for the future compilation of different dictionaries of Slovene, both monolingual and bilingual, and as such its concept is biased towards lexicography. Secondly, it will be used for the enhancement of natural language processing tools for Slovene. The database is organized in six hierarchical levels with lexico-grammatical information which spans from simple morphological data on the top level to semantic, syntactic and collocational data on subordinate levels, with corpus examples at the bottom. Sketch Engine tool with word sketch, tickbox lexicography and GDEX modules is used to enable faster and more efficient extraction of corpus data from the 620-million word FidaPLUS corpus which is used as the source for the data in the database.

## 1 The “Communication in Slovene” project

Slovene Lexical Database is one of the results of the “Communication in Slovene”<sup>1</sup> project which started in 2008 and will end in December 2013. Other results include: (a) natural language processing tools and resources for Slovene: a statistical tagger and parser with a training corpus and an extensive lexicon with information about word inflection and derivation; (b) language data resources: a billion word written corpus and a million word spoken corpus; (c) a study on language teaching practices in Slovene schools which includes the compilation of a corpus of school essays with teachers' revisions and the analysis of common problems in text production, and (d) language description resources which include a manual of style for writers, a pedagogical corpus-based grammar and the lexical database, all of them freely available in an interactive web portal.

A new web format for language data is considered which will incorporate traditional dictionary information on words and word combinations (senses, collocations, examples, grammatical information etc.), visualization of corpus data and semantic ontologies, real-time exploration of web data, question-answering system etc. The portal is intended both for school population and for general use. Information from

---

<sup>1</sup> The operation is partly financed by the European Union, the European Social Fund, and the Ministry of Education and Sport of the Republic of Slovenia. The operation is being carried out within the operational programme Human Resources Development for the period 2007–2013, developmental priorities: improvement of the quality and efficiency of educational and training systems 2007–2013. Project web page: <http://www.slovenscina.eu/>.

the lexical database will be used in two different contexts: (a) together with the lexicon and other resources it will be integrated in the portal for automatic generation of answers to questions such as “how is this word declined/conjugated; what does it mean; how do I spell it”, and (b) it will be used as a lexico-grammatical resource to be used in natural language tools for Slovene.

## 2 The structure of data in the lexical database

With regard to different user needs, there are two types of information in the Slovene lexical database. First, lexico-grammatical information will be used for different functions of the portal and intended for human end users, such as sense descriptions in the s.c. semantic frames, representing the starting point for whole sentence definitions, collocations attributed to particular senses of the lemma, and typical examples from the corpus. Second, different types of information are designed for natural language processing tools. These are encoded in a more complex way and – in addition to their immediate use in NLP tools – need an expert to process or interpret them. Among them are the formal encoding of syntactic patterns on the phrasal and clause level as well as the formal encoding of semantic arguments and their types. The database is conceptualized as a network of interrelated lexico-grammatical information on six hierarchical levels with the semantic level functioning as the organizing level for the subordinate levels.

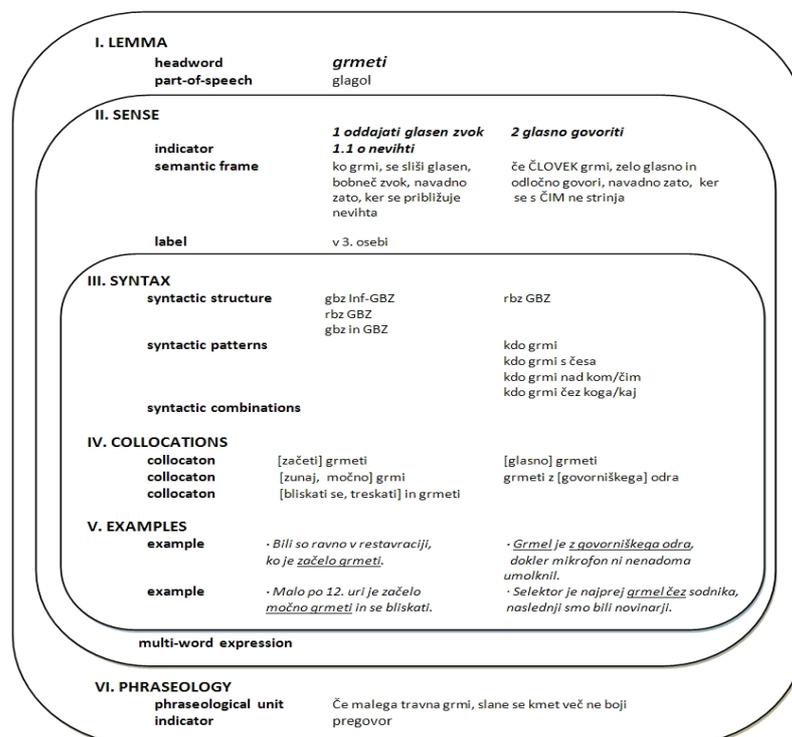


Fig.1. Structure of data in the lexical database

## 2.1 Lemma

Lemma – or the headword – represents the top hierarchical level and functions as the umbrella for all lexical units placed under it: senses and subsenses, multi-word expressions and phraseological units. Multi-word expressions are recorded only for nouns and adjectives and placed within particular senses and subsenses. Phraseological units are recorded in a separate section outside the sense or subsense structure. Both multi-word expressions and phraseological units can be given headword status if they show complex semantic structure and/or high frequency in the corpus. On the lemma level each headword is classified as pertaining to one of four parts-of-speech: noun, verb, adjective or adverb. Function words classes are not recorded in the lexical database as entries. Lemma is considered as linked to its inflectional paradigm in the lexicon, therefore word class conversions (e.g. noun-adjective or vice versa) are analyzed on the sense level and not as new entries in the database.

## 2.2 Sense/subsense

On the sense level, senses and subsenses of the lemma are specified. Therefore, a two-level hierarchy is allowed for with the possible role of the upper level to function as an empty category subsuming the subsenses pertaining to a common semantic field, as in the case of the lemma “grmeti” (to thunder) in Figure 2. All senses and subsenses are labelled with semantic indicators whose primary function is to form a sense menu intended for easy navigation within a polysemic entry structure.

**grmeti** *glagol*

<b>1 oddajati glasen zvok</b> <b>1.1 o nevihti</b> <b>1.2 o napravah</b> <b>1.3 o orožju</b> <b>1.4 o glasbi</b> <b>1.5 padati</b> <b>2 glasno govoriti</b>
---

**Fig.2.** Sense menu of the verb “grmeti” (to thunder)

Another kind of information recorded on the sense level are semantic frames which are conceptually close to frames in the FrameNet project [1] [2] and to prototypical syntagmatic patterns in the Corpus Pattern Analysis system [3]. With verbs, as well as some nouns and adjectives, semantic frames are used to record argument structure and semantic types found in a particular sense or subsense. Therefore, semantic frames provide a link between a particular sense of the headword made explicit by semantic indicators, and syntactic conditions for its realization. At the same time, they represent the starting point for creating the whole-sentence definitions similar to the ones found in Cobuild dictionaries [4] [5].

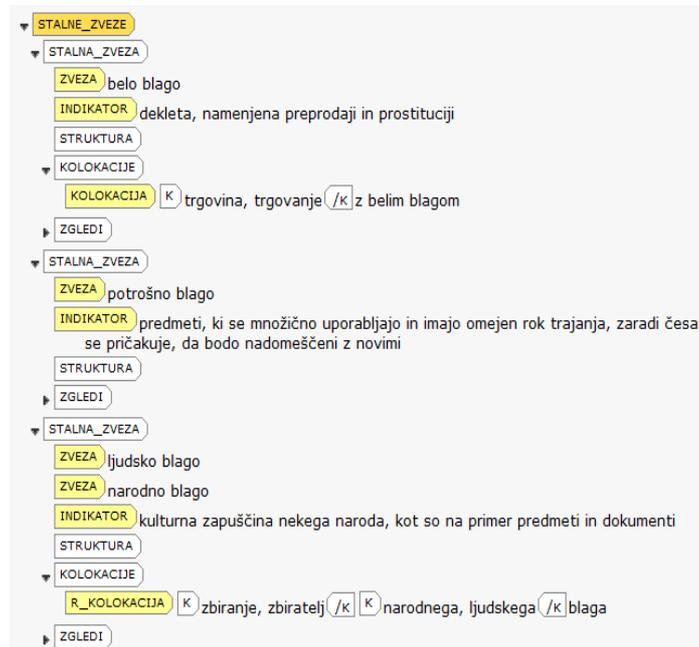
<p><b>sesti</b> <i>glagol</i></p> <p><b>3.2 ustrežati</b> neformalno</p> <p>če neka DEJAVNOST, STANJE ali LASTNOST česa ČLOVEKU sede, mu ugaja ali ustreza</p>	<p><b>nota</b> <i>samostalnik</i></p> <p><b>1 značilnost</b></p> <p>če LASTNOSTI česa dajejo IZDELKU, KRAJU ali DEJANJEM svojo noto, se v njem izražajo in ga delajo posebnega</p>	<p><b>pozoren</b> <i>pridevnik</i></p> <p><b>3 ustrežljiv; obziren; uvideven</b></p> <p>če je ČLOVEK pozoren do drugega ČLOVEKA, je do njega ustrežljiv in skrben ter mu izkazuje pozornost</p>
--	--	---

**Fig.3.** Semantic frames from a verb, noun and adjective entry

Whole-sentence definitions in the form of if-clauses include information about typical syntactic patterns (lemma “aktiven” – active: /predicative use/ a PERSON is active in a FIELD or an ACTIVITY if he/she is participating in it on a regular basis; /attributive use/ an active DEVICE or COMPUTER PROGRAM is ready to function immediately), reflexivity (lemma “briti” – to shave: if a PERSON shaves his/her HAIR or if a PERSON shaves another PERSON's HAIR, he/she removes them with a DEVICE), pragmatic aspects of headword usage (lemma “lahkomiseln” – ≈loose: a WOMAN is considered loose if she likes to be in the company of different men or frequently changes partners), or grammatical limitations (lemma “grmeti” – to thunder: /only in 3<sup>rd</sup> pers. sing./ if it thunders in an AREA or if WEAPONS thunder, shooting can be heard). Semantic types are linked to other kinds of information on subordinate levels thus enabling the user to access data at different level of abstraction, from natural and explicit corpus contexts to implicit semantic types.

### 2.3 Multi-word expressions

Multi-word expressions are registered only in the entries with noun or adjective headwords, either within a particular sense/subsense or after all registered senses and subsenses if semantic relation cannot be established between the MWE and one of the sub/senses. Multi-word expressions must demonstrate a non-compositional idiosyncratic sense, again described by a semantic indicator, mostly identifying the rather broad semantic field or domain. MWUs can show variant forms which are listed under the same MWU entry section and can have their own collocations.



**Fig.4.** Multi-word expressions with other kinds of information in the DPS Entry Editor software

## 2.4 Syntactic structures

Clause patterns: a degree of syntactic information in the form of patterns is already present in the clause structure within semantic frames, e.g. for the verb “sesti” (to sit) in the sense “to agree with, to suit” the pattern “kaj ustreza komu” (sth agrees with sb) is registered, for the noun “nota” (a note) in the sense of “characteristics” the pattern “kaj daje noto čemu” (sth gives a particular note to sth) is registered, and for the adjective “pozoren” in the sense “attentive, tender, caring” the pattern “kdo je pozoren do koga” (sb is attentive to sb) is registered. On syntactic level in LBS, clause patterns are registered systematically with pronouns (sth, sb) in place of semantic arguments to account for alternations of the prototypical pattern in the manner of the theory of norms and exploitations [6]. Alternations include cases where particular arguments are realized by different syntactic possibilities (prepositional phrases, subordinate clauses etc.) or not at all (as in inherent arguments, e.g. in the case of the verb “dihati” – to breathe, “air” is the inherent argument since it rarely expressed as the object: to breathe air). Clause patterns represent useful information for grammar writing and for teaching of Slovene as foreign language, and they enable automated transition from prototypical patterns in semantic frames to typical alternations of these patterns.

**grmeti** *glagol*

**2 glasno govoriti**

če ČLOVEK grmi, zelo glasno in odločno govori, navadno zato, ker se s ČIM ne strinja ali je jezen

a) Struktura:

- ▶ kdo/kaj grmi
- ▶ kdo grmi s česa
- ▶ kdo grmi nad kom/čim
- ▶ kdo grmi na koga/kaj
- ▶ kdo grmi čez koga/kaj
- ▶ kdo grmi zoper koga/kaj
- ▶ kdo grmi proti komu/čemu
- ▶ kdo grmi o čem
- ▶ kaj grmi od česa

**Fig.5.** Clause patterns of the verb “grmeti” in the sense “to talk loudly”

Syntactic structures and collocations: syntactic structures represent a formalization of typical patterns on the clause and phrasal level and are primarily intended for natural language processing tools. They are registered in the form of syntagmatic combinations of words and phrases, and are composed of a part-of-speech label plus the information on grammatical case. Formally, labels are conformant with morpho-syntactic tags used in the FidaPLUS/Gigafida corpus [7] in the Sketch Engine tool [8]. Listed types of syntactic structures are partly dependant on the part-of-speech of the headword. Within a particular structure, the position of the headword is indicated by the capitalization of the label. Where typical collocates realizing a collocation for a particular syntactic structure exist, they are registered under that structure. With verbs, syntactic structures include verbal phrases with infinitives, adverbs or coordinating structures among others, as shown in Figure 6.

**grmeti** *glagol*

**2 glasno govoriti**

b) Struktura: rbz GBZ

- [glasno] grmeti

c) Struktura: gbz Inf-GBZ

- [začeti] grmeti

d) Struktura: gbz in GBZ

- [bliskati se, bliskati] in grmeti

**Fig.6.** Syntactic structures and collocations of the verb “grmeti” in the sense “to talk loudly”

The source for extracting syntactic structures and collocations from the corpus are s.c. word sketches in the Sketch Engine tool described below. The number of syntactic structures is finite – at the time of writing almost 300 structures are recorded, however, not all of them exhibit collocations, as in the case of the structure /pbz0 SBZ0/ (adjectival phrase + nominal phrase, both with the whole inflectional paradigm) shown in Figure 6.

## 2.5 Syntactic combinations

Syntactic combinations represent an intermediate level between collocations and multi-word expressions. The most typical members of the class are prepositional phrases and other multi-word combinations which extend beyond the binary syntactic structures, but are on the other hand compositionally fixed and have at least one invariable lexical element, together with another lexically variable but syntactically obligatory element. Syntactic combinations also include elements with numerical expressions, comparisons and coordinate structures. Contrary to multi-word expressions, syntactic combinations do not need explanation in semantic terms (their meaning is compositional) and therefore no indicators or semantic frames are provided. Consequently, they cannot be given headword status.

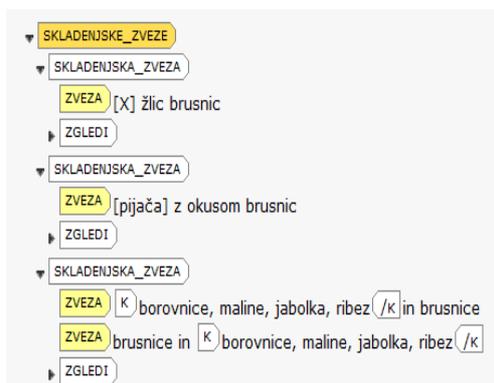


Fig.7. Syntactic combinations under the headword “brusnica” (cranberry) in the “fruit” sense

## 2.5 Collocations and examples

On the collocation level, patterns and structures are verified by recording typical collocates of the headword realized in the anticipated syntactic positions. At the same time, higher levels in the hierarchy are taken into account and the same collocation can be recorded under different syntactic structures or senses if it shows semantic diversity. On the last level in the hierarchy, collocations and also all parent levels (patterns, structures and frames with semantic types) are verified by recording corpus examples using GDEX [9] and TBL tools [10] in the Sketch Engine.

### 3 Corpus data and tools

#### 3.1 The corpus

At the time of writing, corpus used for the compilation of the lexical database is the FidaPLUS corpus [11] with 620-million words and containing text from 1990–2006. In 2011 it will be replaced by another generation of the corpus line which started in 2000 with the 100-million word FIDA corpus with restricted access, which was upgraded and made publicly available as FidaPLUS corpus in 2006. The new corpus is called Gigafida and contains 1.1 billion words from texts of different genres spanning from 1990–2010. Its composition and characteristics are described in [7]. Together with the new text data, a new web interface was developed with a particular focus on user friendliness and ease of access to the data for non-expert users. However, for the purpose of lexical database compilation, the FidaPLUS corpus was put into the Sketch Engine tool and this version of the corpus is used by the lexicographers, making use of advanced corpus query features provided by the tool.

#### 3.2 Sketch Engine

Sketch Engine represents the basic lexicographic corpus data extraction tool used by the lexicographers compiling Slovene lexical database. Together with the standard use of a concordances with advanced options such as the use of corpus query language (CQL) and similar, two additional features are used which enable faster compilation of the database. The first is word sketches module which provides one-page automatic summaries of a word's grammatical and collocational behaviour. Word sketches are based on s.c. sketch grammar where grammatical relations are defined as regular expression over POS-tags. Slovene sketch grammar currently contains 32 grammatical relations or gramrels which basically reflect the 300 recorded syntactic structures. The other feature are the combined Tickbox lexicography and GDEX modules which provide a faster way to select good dictionary examples recorded under each structure and collocation in the database. The module described in [9] and [10] was upgraded for its use with the Slovene language.

### 4 Conclusions

The concept of the Slovene lexical database is biased towards lexicography but its intended use is also the enhancement of natural language processing tools for Slovene, such as taggers and parsers. Along with the use of data within the language portal for human end users, we believe that syntactic structures and patterns recorded in the database will make contribution to the better quality of the parser which is also under development in the same project. This assumption will be tested at the end of the compilation process when automated extraction of data from the corpus will be tested. Further research is foreseen also with the analysis of semantic type recorded in semantic frames where an ontology could be constructed and linked to the FrameNet and/or WordNet data. Lastly, we expect that the database will be used for the first automatic word sense disambiguation experiments for the Slovene language.

## References

- [1] Fillmore, Ch. J., Atkins B.T.S. (1992). Towards a frame-based organization of the lexicon: the semantics of RISK and its neighbors. In: Lehrer, A., Feder K.E. (eds.) *Frames, Fields, and Contrasts*. Lawrence Erlbaum Associates, 75-102.
- [2] Baker, C.F., Fillmore, C.J., Cronin, B. (2003). The Structure of the Framenet Database. *International Journal of Lexicography*, 16(3), 281-296.
- [3] Hanks, P. (2004). Corpus Pattern Analysis. In: Williams, G., Vessier, S. (eds.) EURALEX 2004. *Proceedings*. Lorient: Université de Bretagne-Sud.
- [4] Barnbrook, G., Sinclair, J. (1994). Parsing Cobuild Entries. In: Sinclair, J., Hoelter, M., Peters, C. (eds.) *The Languages of Definition: The Formalisation of Dictionary Definitions for Natural Language Processing*. Luxembourg. European Commission. 13-58.
- [5] Barnbrook, G. (2002). *Defining Language: A Local Grammar of Definition Sentences*. Studies in Corpus Linguistics: John Benjamins Publishing Company.
- [6] Hanks, P. (1994). Linguistic Norms and Pragmatic Exploitations, or Why Lexicographers need Prototype Theory and Vice Versa. In: F. Kiefer, G. Kiss, and J. Pajzs (eds.) *Papers in Computational Lexicography: Complex '94*. Research Institute for Linguistics, Hungarian Academy of Sciences.
- [7] Logar Berginc, N., Krek, S. (2010) New Slovene corpora within the Communication in Slovene project. International Conference SLAVICORP. Corpora of Slavic Languages. 22-24 November 2010. *Abstract*, 8.
- [8] Kilgarriff, A., Tugwell, D., (2001). WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. *Proceedings of the AVL workshop on COLLOCATION: Computational Extraction, analysis and Exploitation*. Toulouse. 32-28.
- [9] Kilgarriff, A., Husak, M., McAdam, K., Rundell, M., Rychly, P., (2008). Automatically finding good dictionary examples in a corpus. In: Bernal, E., DeCesaris, J. (eds.) *Proceedings of the XIII EURALEX International Congress. Barcelona, 15-19 July 2008*. Barcelona: Documenta Universitaria: Institut universitari de lingüística aplicada, Universitat Pompeu Fabra, 425-432.
- [10] Kilgarriff, A., Kovar, V. & Rychlý, P. (2009). Tickbox Lexicography. In: Granger, S. & Paquot, M. (eds.) *Proceedings of eLex2009: eLexicography in the 21st Century: New Challenges, New Applications*. Louvain-la-Neuve, Belgium. 411-418.
- [11] Arhar, Š., Gorjanc, V., Krek, S. (2007). FidaPLUS corpus of Slovenian: the new generation of the Slovenian reference corpus: its design and tools. In: *Proceedings of the Corpus Linguistics Conference, CL2007, University of Birmingham, UK, 27-30 July 2007*. Birmingham, 2007.

# Building Annotated Corpora without Experts

Marek Grác

Faculty of Informatics, Masaryk University, Brno

**Abstract.** In this paper, we present a low-cost approach of building a multi-purpose language resource for Czech, based on currently available results of previous work done by various teams. We focus on the first phase that consists of verifying validity of automatically discovered syntactic elements in 10 000 sentences by 47 human annotators. Due to the number of annotators and very limited time for training, existing heavy-weight techniques for building annotated corpora were not applicable. We have decided to avoid using experts when results between annotators differed. This means that our corpus does not offer ultimate answers, but raw data and models for obtaining “correct” answer tailored to user’s application. Finally we discuss the currently achieved results and future plans.

## 1 Introduction and motivation

Language resources for natural language processing are very important for development as well as improvement of existing natural language processing (NLP) tools. Situation for different European languages varies a lot. In the worst case there are almost no resources and we face the problem of creating them in a cheaply and quickly while maintaining high quality. Different NLP applications require both different data and different annotations. We can attempt to build an ultimate corpus that will be useful for every application but we do not believe that such approach is successful often enough. We have decided to model our corpus using application-driven development. This approach should prevent major design flaws which might not be automatically recoverable later and could limit the usefulness of resulting work for our needs.

Using this approach, we have decided to build a multi-layered annotated corpus where each layer has to be created for a specific set of applications. We avoid creating data with no immediate application, even if it might be useful in the future. This kind of development helps us to create usable data (and useful in our applications) right after the annotation of any given layer is finished.

Czech language is one of the most described European languages and there are several high quality resources like PDT2 ([2]) or Czech version of EuroWordNet ([10]). We chose to use morphologically annotated corpus DESAM [9], that was manually checked by annotators and contains 50 000 sentences. Due to lack of resources we took only first 10 000 sentences from this corpus but we plan to expand our coverage after we will verify that cost of annotating more data gives us enough additional value.

In the first phase (described in this article) we want to perform computer-aided annotation of syntactic elements (that will be described further on). Second phase (not yet finished) is dedicated to annotating relations between identified syntactic elements. These results will show us PP-attachment and relations between NP and verb, thus creating structure very similar to simplified predicate-argument structure. In the third phase we plan to

semi-automatically match previous layer with valency frames from VerbaLex ([4]) and create an ontology based on valency frames.

## 2 Annotation process

Annotation of linguistic data is considered to be a task for experts. This is especially right for those corpora that attempt to cover more layers or structures of a language. Process of annotation is usually described in detail in an annotation manual. As an example, we can take annotation manual for the syntactic layer of PDT2.0 which spans 301 pages [3]. In last years, we have witnessed several attempts to use crowdsourcing for small parts of linguistic annotation [8]. However, we are not aware of any attempts to build annotated corpora completely via crowdsourcing. In order to use crowdsourcing we have to find a crowd that exceeds a critical mass. Thanks to services like Amazon Mechanical Turk, this is usually not a problem for widely used languages, such as English. Situation for languages like Czech (10 million speakers) is more complicated as no services of this type are available.

We have decided to involve students. Our annotators were mostly in their first year at the university and they have very limited amount of deeper linguistic knowledge. Our previous experience with student annotators gives us some hope that they can be trained to carry out simple linguistic tasks better than an average crowd-member, though.

We assume that an annotation standard is usually an attempt to approximate several mutually exclusive and contradictory constraints [5]:

1. **completeness**: the annotation should provide complete linguistic insight into the particular area;
2. **consistency**: the annotation should be consistent, i. e. same or similar language phenomena should be handled in same or similar ways;
3. **usability**: the annotation should enable straightforward usage in the intended applications;
4. **simplicity**: the annotation should be as simple as possible to make high inter-annotator agreement achievable.

In our experience, most language resources try to find a trade-off among the constraints by prioritizing them in the order given above. They prefer completeness over consistency, and both of them over simplicity.

Following the so-called KISS<sup>1</sup> principle, we are strongly convinced that the reverse order of those constraints represents a much better priority list to be met when building a language resource. Thus, our priorities are:

- **simplicity**: so that annotators do not err too often;
- **usability**: so that the usage of the resource will be straightforward;
- **consistency**: following from simplicity;
- **completeness**: just in case everything is simple, usable and consistent.

Main objection against this new order of priorities can be that consistency is crucial to most NLP application. This applies to using the data both for testing/development and

<sup>1</sup> Keep It Simple Stupid

for machine learning. From our perspective, natural language is too ambiguous and flexible to be easily and consistently annotated. We have to face situations where even expert human annotators encounter a possibility of having more than one correct annotation. Inconsistencies between annotators are traditionally resolved by an expert who decides which annotation is correct. Qualified opinion of an expert can improve consistency of annotations. Our approach to building corpora does not call for any experts, so we have to cope with that problem in a different way. In our approach, inconsistencies between annotators are instead resolved by a resolution model that gives us the desired results. We offer all annotations together with annotator experience to the developer so he can choose the correct model for his application. For some applications, we find it correct to use all results which are confirmed by 3 of 4 annotators but for other purposes we want 100% agreement. Resolution model helps us to select a trade-off between the amount of annotated data that fulfill our criteria and their consistency.

Simplicity is at the top of priorities and we plan to reach that goal with constraining the annotators as much as possible with a simple annotation scheme. In the first phase, annotators can answer only yes/no and they have no possibility to add a correct solution even if they believe that they know it. Limiting creativity and working with preprocessed data helps us to increase inter-annotator agreement and (therefore) also consistency.

In the first phase of building an annotated corpus that is described in this article, we have 47 annotators and majority of them worked on the annotations only for 4 or 8 hours. Therefore, using a detailed annotated manual is not a viable option – the annotators would spend more time studying the manual than annotating data. Under these constraints, we have decided to give them only a very short annotation manual and a set of examples, letting them decide what is right based on their natural understanding of the language.

### 3 Syntactic elements annotation

The first phase of annotation is to identify syntactic elements in corpora. We have tried to identify clauses, verb phrases, noun phrases and coordination of noun phrases. These types are often used in shallow syntactic analysis and with the exception of noun phrases we follow the general definitions. Situation with noun phrases is slightly more complicated, so we will explain it more precisely in this section.

We soon realized that we have to distinguish between two basic types of noun phrases. Their precise description can be found in the next section; for now, an example is enough for us. We will borrow one of the famous examples – ‘I saw a man with a telescope’.

There are two short (minimal) noun phrases ‘a man’ and ‘a telescope’ and we do not attempt to infer their relations. We could find a *maximal* noun phrase (‘a man with a telescope’) but maximal noun phrases are determined semantically and we cannot expect high inter-annotator agreement for them. Minimal noun phrases can be described more precisely and agreement between annotators is expected to be much higher. Even those minimal phrases can be very useful for various applications, including syntactic analysis. We do not plan to forfeit maximal noun phrases, however. We will create them from existing minimal phrases in next phase of our project. This will provide us with additional quality checks for the minimal noun phrases and thus lead to better consistency in later phases.

We have obtained first 10 000 sentences from morphologically desambiguated corpus of Czech language named DESAM. These sentences were parsed by a rule-based syntactic

analyser, SET ([6]), and syntactic structures VP, NP, CLAUSE, COORDINATION were extracted. Annotators had no option to enter a valid phrase if it was not identified by the automatic analyser. As we can see, our corpus is not going to contain syntactic structures which we cannot find automatically. Nevertheless, we are adding new results from syntactic analysers to the corpus, so its coverage is increasing. Our situation is very close to real world as we cannot expect that our analysis will be completely correct and very usually we are working with only partial information.

Annotators received a block of 400 sentences for which they had to classify identified elements as valid or invalid. They were not aware that we will not ask experts to evaluate their work and that we only have very limited possibilities to check the quality of their work. As a basic metric for their evaluation, we chose the ratio of accepted/rejected syntactic elements, inter-annotator agreement with different annotator working on same block and the number of syntactic elements which were left unmarked. Only one of the annotators had problems to fulfill our criteria and she had to annotate the same block again. These are very good results considering that the annotators received about 15 minutes of training and a two-page annotation manual.

#### **4 Technical background and annotation tool**

Corpus sholva will contain several layers of linguistic annotation and for this reason, we have decided to use a NXT NITE toolkit ([11]) which was developed for multimodal corpora. We do not plan to build a multimodal corpus, but existing libraries for searching through relations between elements in corpora and the XML format persuaded us. On top of this toolkit, we have built our own library, libSholva, which maps elements in the corpus into objects, so that programmer does not need to care about internal NXT NITE structures or about XML elements. On top of this stack, we have developed a graphical user interface for annotators called pHrase annotator. Annotators were able to complete work in it without any further training and they were able to classify around 100 sentences per hour. We can compare these numbers with complete linguistic annotation by inexperienced annotators in PDT2, who were able to annotate 8 sentences per hour on average [7]. It is clear that an annotator doing full linguistic annotation creates a resource that can be reused in other types of application, but we would like to stress that this traditional approach cannot be applied to produce useful results with only limited resources: we need at least thousands of sentences for even very basic testing.

#### **5 Obtaining correct information from corpus**

Attempts to annotate most of a linguistic resource have to cope with the cases where two annotators are of a different opinion. This can happen due to an incomplete annotation manual, mis-clicking, fatigue or the fact that language is ambiguous. Corpus that wants to provide only correct answers will usually call an expert to decide which opinion is correct and when both opinions are, find the “better” opinion. It is very important to identify these cases and be consistent in their treatment. This is quite complicated in the cases of actual language ambiguity, when looking for the “better” opinion, since it is very difficult to formalise the reasoning behind such a decision.

Currently, we use three basic models for obtaining “correct” result:

- all annotators have to agree
- one annotator may disagree (for problematic cases there are usually at least 4 independent annotations)
- weighted generalized model where weights are selected according to annotator experience.

This can help us to find a proper result depending on application. If we find invalid cases automatically then we can add a new “computer” annotator with bigger weight to overrule previous results. However, we are not removing any annotations from the corpus.

## 6 Annotation results

When using non-experts annotators, it is very important to find the right interpretation of the results. We are applying described resolution models to raw data on a per-application basis, but those models only help us to get the best of the raw data – they have no impact on the quality of the raw data.

Since we have no gold standard, we have to compare quality of annotation between the annotators themselves. If the inter-annotator agreement is high, we can assume that annotators are able to work consistently and according to the annotation manual and their linguistic knowledge. For measuring inter-annotator agreement, we can use several methods. We have chosen three of them.

The first method is based on the ratio of syntactic elements marked as valid or invalid. After annotating a substantial amount of corpus, we were able to identify the range of expected values (0.7 – 0.9). If annotators did not fit into this range, the results were not accepted. Only one of the annotators failed to fall within this range. Therefore, we assume that the annotators were able to understand their task and work on it in a similar manner.

Second method can be used for comparison between annotators who have worked on the same data block (400 sentences). We identify elements which were marked differently by different annotators and since those results are not reliable yet they have to be confirmed by other annotators. This absolute measurement was used for creating a set of unclear syntactic elements and those results were re-annotated by two additional annotators. The main problem of this measurement is that even if we can obtain high numbers (>65%), we cannot prove that they are significantly better than agreement by chance. This is a problem since the annotator can choose only between two answers and we know from the ratio between valid/invalid elements that we can obtain agreement of more than 65% from purely random answers.

Problems of agreement by chance are especially pertinent when we are using Cohen’s kappa [1] to measure inter-annotator agreement. This coefficient takes into account the number of possible answers and frequency of their usage. Cohen’s kappa can be applied to evaluate agreement between two or more annotators. For illustration purposes, we have decided to compute it only for pairs of annotators as these are the numbers that are usually published and can therefore be compared readily. Interpreting the obtained numbers can be difficult for those who are unfamiliar with Cohen’s kappa. For this reason, several benchmark scales were developed (table 6). Our first results were very close to the border between “moderate” and “substantial”, according to Landis-Koch, but in the later stages, more experienced annotators were able to reach upper limit of “substantial” agreement. Since most of our elements are valid, we cannot realistically expect to obtain numbers in

the “almost perfect” range. In either case, this means that our results are much better than coincidence.

Landis and Koch		Fleiss	
Kappa Statistics	Strength of Agreement	Kappa Statistics	Strength of Agreement
< 0.0	Poor		
0.0 to 0.20	Slight		
0.21 to 0.40	Fair	< 0.40	Poor
0.41 to 0.61	Moderate		
0.61 to 0.80	Substantial	0.40 to 0.75	Intermediate to Good
0.81 to 1.00	Almost Perfect	0.76 to 1.00	Excellent

**Table 1.** Example of Benchmark Scales for Cohen’s Kappa

Since we would like to strengthen our belief of better than random results, we have decided to introduce noise into the data. The grammar of our syntactic analyser was changed to randomly prefer invalid noun phrases (1.5%). Those noun phrases were slightly special, since they did not contain all words between the first and the last word. Such phrases can exist in Czech but usually they are not the minimal ones. Annotators tend to mark them as valid in range (0% - 6%). For our purpose, it is very important that annotators did not mark those as valid and no such noun phrase was identified by more than one annotator.

valid by annotator A/B	invalid by annotator A/B	agreement	cohen kappa	max kappa
2377/2604	975/793	79.04%	0.50	0.85
3009/2669	326/685	88.43%	0.66	0.81
2515/2664	754/606	92.05%	0.73	0.86
2473/2431	678/713	92.68%	0.76	0.96

**Table 2.** Annotation results

Table 6 contains both the best and the worst result of inter annotator agreement. In the case of first line, the concept of verb phrases was not understood correctly by one of the annotators and there were 3 – 5 times as many errors as usual. All the other elements were in the usual range.

## 7 Conclusion

In the paper, we have presented a concept of application-driven development of a multi-purpose linguistic resource. We have described the first phase of the project, which consists of computer-aided annotation of syntactic elements: clauses, verb phrases, minimal noun and prepositional phrases and coordination of noun phrases. We also displayed some results of inter-annotator agreement on validating these elements.

In the future, the project will continue with the annotation of additional layers. We plan to release our corpus to the research community after the second phase of the project

(identifying relations between elements) will be completed. Along with that, our linguistic tools and resources will be improved by fixing problems discovered in the process of annotation.

The infrastructure for building this type of annotation corpus is freely available at <http://www.bushbank.org> and we will gladly help to create a similar resource for other languages. We believe that this can be a way for even smaller languages to obtain valuable linguistic resources, using a very low-cost approach.

## Acknowledgments

This work has been partly supported by the Academy of Sciences of Czech Republic under the project T100300419, by the Ministry of Education of CR within the Center of basic research LC536.

## References

- [1] Cohen, J. (1960). A coefficient of agreement for nominal scales. In *Educational and Psychological Measurement*, volume 20, pages 37–46.
- [2] Hajič, J. (2004). Complex Corpus Annotation: The Prague Dependency Treebank. Bratislava, Slovakia. Jazykovedný ústav Ľ. Štúra, SAV.
- [3] Hajič, J., Panevová, J., Buráňová, E., Urešová, Z., Štěpánek, J., Pajas, P., and Kárník, J. (2005). Anotace na analytické rovině – Návod pro anotátory.
- [4] Horák, A. and Pala, K. (2007). Building a large lexicon of complex valency frames. In *Proceedings of the FRAME 2007: Building Frame Semantics Resources for Scandinavian and Baltic Languages*, pages 31–38, Lund University, Sweden. Tartu, Estonia.
- [5] Jakubíček, M., Kovář, V., and Grác, M. (2010). Through low-cost annotation to reliable parsing evaluation. In *PACLIC 24 Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*.
- [6] Kovář, V., Horák, A., and Jakubíček, M. (2009). Syntactic analysis as pattern matching: The SET parsing system. In *Proceedings of the 4th Language & Technology Conference*, pages 100–104, Poznań, Poland.
- [7] Mikulová, M. and Štěpánek, J. (2010). Ways of evaluation of the annotators in building the prague czech-english dependency treebank. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 1836–1839, Valletta, Malta. European Language Resources Association.
- [8] Munro, R., Bethard, S., Kuperman, V., Lai, V. T., Melnick, R., Potts, C., Schnoebelen, T., and Tily, H. (2010). Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk, CSLDAMT ’10*, pages 122–130, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [9] Pala, K., Rychlý, P., and Smrž, P. (1997). DESAM – annotated corpus for Czech. In *Proceedings of SOFSEM’97*, pages 523–530. Springer-Verlag. Lecture Notes in Computer Science 1338.
- [10] Pala, K. and Smrž, P. (2004). Building Czech WordNet. *Romanian Journal of Information Science and Technology*, 7(2-3):79–88.

- [11] Soria, C., Bernsen, N. O., Cadee, N., Carletta, J., Dybkjaer, L., Evert, S., Heid, U., Isard, A., Kolodnytsky, M., Lauer, C. and Lezius, W., Noldus, L., Pirrelli, V., and Reithinger, N. (2002). Advanced tools for the study of natural interactivity. In *Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Spain.

# The Lexicographic Representation of Czech Diatheses: Rule Based Approach

Václava Kettnerová and Markéta Lopatková\*

Charles University in Prague, Faculty of Mathematics and Physics  
Institute of Formal and Applied Linguistics

**Abstract.** In this paper, we propose a method of the representation of Czech diatheses in the valency lexicon of Czech verbs, VALLEX. Under the term diatheses, specific relations between uses of the same verb lexeme are considered here. These relations are associated with changes in valency frames of verbs which stem from the changes in the linking of situational participants, valency complementations and surface syntactic positions. We distinguish three types of Czech diatheses according to which linguistic means they are based on: (i) grammatical, (ii) syntactic and (iii) semantic diatheses. We demonstrate that in case of grammatical and syntactic diatheses, the changes in valency structure of verbs are regular enough to be captured by formal syntactic rules whereas the changes associated with semantic diatheses can be represented rather by lexical rules. In conclusion, we show that on certain conditions the different types of diatheses can be combined together.

## 1 Introduction

Although diatheses have been widely debated in the literature in the past decades, the results of the theoretical research have not been fully applied in the available lexical resources so far. Individual theories dealing with these phenomena are summarized, e.g., in [5]. Here we propose a rule based approach to Czech diatheses for their representation in the valency lexicon of Czech verbs, VALLEX.<sup>1</sup>

Diatheses are defined here as specific relations between uses of the same verb lexeme: these uses exhibit semantic affinity, however, they are syntactically structured in different ways. A question arises how it is possible to describe changes in valency structure of verbs associated with diatheses in the lexicon. When describing these changes, the distinction between situational content and structural meaning plays a key role (Section 2).

In principle, we distinguish three types of Czech diatheses according to which linguistic means they are expressed by: (i) grammatical diatheses (Section 3), (ii) syntactic diathesis (Section 4) and (iii) semantic diatheses (Section 5). This differentiation is warranted by the fact that on certain conditions, the different types of diatheses can be combined together (Section 6).

As to the representation of diatheses, we demonstrate that whereas grammatical and syntactic diatheses can be represented by formal syntactic rules, semantic diatheses require to be described on the basis of lexical rules.

\* The research reported in this paper was carried out under the project of MŠMT ČR No. MSM0021620838. It was supported by the grant No. LC536 and partially by the grants No. GA P406/2010/0875 and P202/10/1333.

<sup>1</sup> <http://ufa1.mff.cuni.cz/vallex/2.5/>

This paper follows and further develops the issues addressed in [5]. In contrast to [5], the previous typology of Czech diatheses is enriched with syntactic diathesis here. Furthermore, an adequate representation of semantic diatheses (introduced in [7] and [6]) is recapitulated. Moreover, the combination of the different types of diatheses and its representation are newly discussed.

## 2 Situational content and structural meaning

In prototypical cases, a single meaning corresponds to a single valency structure. However, in many cases a similar meaning can be syntactically structured in a different way. See the following uses of the verb *semlít* ‘to grind’:

- (1) a. *The millers ground wheat into flour.* – b. *Wheat was ground into flour (by the millers).*  
 (2) a. *The millers ground flour out of wheat.* – b. *Flour was ground out of wheat (by the millers).*

We refer to the specific relations between such uses of a verb as diatheses. For the purpose of their description, we distinguish between a so called situational content of a verb and its structural meaning.

The term *situational content* refers to the lexical-semantic characteristics of a verb which is related to a situation portrayed by the verb. This situation consists of a set of *situational participants* characterized by particular semantic properties and related by certain relations. The situational content represents a syntactically unstructured part of the verbal characteristics.

A syntactically structured part, i.e., such part in which the components of the situational content are syntactically expressed, is referred here as a *structural meaning*. In the Functional Generative Description (henceforth FGD, see esp. [16]), which serves as a theoretical background for VALLEX, the structural meaning corresponds to the tectogrammatical layer, i.e., the layer of linguistically structured meaning. The structural meaning of verb is described by a set of *valency complementations* labeled by tectogrammatical roles [9].<sup>2</sup>

We assume that each lexical unit of a verb is characterized by both situational content and structural meaning in a unique way: a particular set of situational participant(s) which are mapped onto a set of valency complementation(s) is characteristic of each lexical unit. Any changes in the situational content or the structural meaning lead to the change of a lexical unit of the verb.

Diatheses represent such relations between uses of a verb which are characterized by closely related (or the same) situational contents whose situational participants are mapped onto surface syntactic positions in a different way. The changes in the correspondence between situational participants and surface syntactic positions always affect the prominent positions of subject or direct object. Then the situation denoted by a verb is perspectivized in light of the situational participant which is mapped onto the subject or direct object.

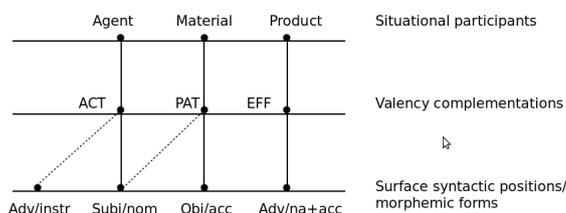
<sup>2</sup> The distinction of the situational content and the structural meaning is inspired by [10] and [8]. Whereas the situational content is taken as a part of the verbal meaning in [10] and [8], here in accordance with FGD [16], syntactically unstructured components of the verbal meaning is taken as content issues.

### 3 Grammatical diatheses

Grammatical diatheses and their representation are extensively discussed in [5] and [7]. Let us recapitulate their brief description. We consider grammatical diathesis as a relation between uses of a verb which are characterized by changes in the mapping between valency complementations and surface syntactic positions while the correspondence between situational participants and valency complementations is preserved. These changes arise from the use of a specific grammatical meaning of a verb and are associated with the shift of ‘Agent/Bearer’ from the prominent surface syntactic position of subject. We illustrate the relation of grammatical diatheses by examples (3a)-(3b):

- (3) a. *Mlynáři*.ACT-Agent *semleli zrna*.PAT-Material *na mouku*.EFF-Product –  
 b. (*Mlynáři*.ACT-Agent) *bylo zrna*.PAT-Material *semleto na mouku*.EFF-Product  
 Eng. a. The millers.ACT-Agent ground wheat.PAT-Material into flour.EFF-Product –  
 b. Wheat.PAT-Material was ground into flour.EFF-Product (by the millers).ACT-Agent

We observe that the uses of the verb *semlít* ‘to grind’ in examples (3a)-(3b) share the same situational content as well as structural meaning. As a consequence, we consider the pairs of the uses of the verb *semlít* ‘to grind’ in (3a)-(3b) as a single lexical unit. Then what differs these uses in (3a)-(3b) is the mapping of the valency complementations onto surface syntactic positions: (i) ‘ACTor’ (in both cases corresponding to the situational participant ‘Agent’) is mapped either onto the prominent surface syntactic position of subject, (3a), or onto a less prominent adverbial position, (3b), and (ii) ‘PATient’ (expressing the situational participant ‘Material’ in both sentences) corresponds either to the position of direct object, (3a), or to the subject, (3b), see Figure 1.



**Fig. 1.** The changes in the mapping of the valency complementations and the surface syntactic positions of the verb ‘to grind’ associated with the passive grammatical diathesis.

The changes in the mapping of the valency complementations and the surface syntactic positions result from the use of a specific grammatical meaning (passive meaning) of the verb *semlít* ‘to grind’. We consider the use of the verb characterized by this grammatical meaning as a marked one. In FGD, these grammatical meanings of verbs are represented by a set of verbal grammatemes, see esp. [13] and [14].

We distinguish the following types of Czech grammatical diatheses. The grammatical meanings of the verbs in the marked constructions of diatheses are captured by the following values of the grammateme *diatgram*: *act* (4a), (5a), (6a), (7a) and (8a), *pass* (4b), *deagent* (5b), *rez1* (6b), *rez2* (6c), *recip* (7b) and *disp* (8b):

- (4) **Passive diathesis**  
 a. *Karel IV. založil<sub>act</sub> roku 1348 Karlovu univerzitu.* – b. *Karlova univerzita byla založena<sub>pass</sub> v roce 1348 (Karlem IV.).*  
 Eng. a. Charles IV. founded<sub>act</sub> Charles University in 1348. – b. Charles University was founded<sub>pass</sub> (by Charles IV.) in 1348.
- (5) **Deagentive diathesis**  
 a. *Snídani podáváme<sub>act</sub> mezi sedmou a devátou hodinou.* – b. *Snídaně se podává<sub>deagent</sub> mezi sedmou a devátou hodinou.*  
 Eng. a. We serve<sub>act</sub> breakfast between 7am and 9am. – b. ‘Breakfast – refl – serves<sub>deagent</sub> – between 7am and 9am.’
- (6) **Resultative diathesis**  
 a. *Sekretářka mi zde objednala<sub>act</sub> stůl pro čtyři osoby.* – b. *Je zde objednan<sub>rez1</sub> stůl pro čtyři osoby.* – c. *Mám zde objednan<sub>rez2</sub> stůl pro čtyři osoby.*  
 Eng. a. My secretary booked<sub>act</sub> a table for four persons here. – b. There is booked<sub>rez1</sub> a table for four persons. – c. ‘Have – here – booked<sub>rez2</sub> – a table – for four persons.’
- (7) **Recipient passive diathesis**  
 a. *Šéf mi přidělil<sub>act</sub> novou pracovní.* – b. *Dostal jsem přidělen<sub>recip</sub> novou pracovní (od šéfa).*  
 Eng. a. My boss allocated<sub>act</sub> me a new study. – b. ‘Gave – allocated<sub>recip</sub> – a new study – (by my boss).’
- (8) **Dispositional diathesis**  
 a. *Čtu<sub>act</sub> tento překlad.* – b. *Tento překlad se mi čte<sub>disp</sub> dobře.*  
 Eng. a. I read<sub>act</sub> this translation. – b. This translation reads<sub>disp</sub> well.

**The representation of grammatical diatheses.** We observe that in case of grammatical diatheses the changes in a valency frame of a verb are limited only to changes in morphemic forms of valency complementations. These changes are regular enough to be captured by formal syntactic rules. These syntactic rules are stored in the grammar component of the lexicon. In the data component, there is a single lexical unit representing both uses of a verb. This lexical unit is characterized by a valency frame corresponding to the unmarked use. The possibility of applying some of the rules is ascribed to each relevant lexical unit of a verb. Let us illustrate these syntactic rules by a rule for passive diathesis deriving the marked use of the verb *semlít* ‘to grind’ in (3b), see Table 1.

Pass.r	Unmarked	Marked
verbal grammateme diatgram	act	pass
valency frame	ACT <sub>nom</sub> PAT <sub>acc</sub>	ACT <sub>instr</sub> PAT <sub>nom</sub>

**Table 1.** Pass.r rule for the passive diathesis.

The change in the grammatical meaning of a verb is represented by the change of the value of the verbal grammateme diatgram which is changed from act into pass. Furthermore, the rule describes the changes of morphemic forms of the valency complementations ‘ACTor’ and ‘PATient’. These changes manifest their surface syntactic shift: (i) the shift of ‘ACTor’ from the subject into the adverbial position is expressed by the change of its morphemic form from nominative into instrumental and (ii) the change of

surface syntactic expression of ‘PATient’ is captured by the change of its morphemic form from accusative into nominative (more detailed description of the rule is provided in [5]).

For the representation of the other above listed types of grammatical diatheses, other syntactic rules are formulated. On their basis, the marked members of grammatical diatheses can be derived from each lexical unit of a verb to which they are assigned.

#### 4 Syntactic diathesis

Syntactic diathesis is related esp. to a reciprocity in Czech. It represents another type of the relation between different uses of a verb which are characterized by the changes in the correspondence of valency complementations and surface syntactic positions while the linking of situational participants and valency complementations remains unchanged. For these reasons, similarly as in the case of grammatical diatheses, we consider the members of syntactic diathesis as two surface syntactic expressions of a single lexical unit of a verb. However, in contrast to grammatical diatheses, syntactic diathesis is not underlain by the use of any specific grammatical meaning, i.e., the grammatical meaning of a verb is preserved.

In Czech, the marked members of syntactic diathesis are represented by reciprocal constructions which result from reciprocalization, i.e., a syntactic operation on valency frames of verbs in which two (or three) valency slots – if their features allow for symmetrical usage – are used reciprocally, see esp. [11] and [12].

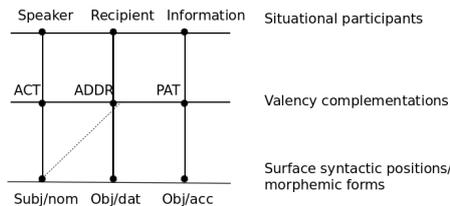
As to the changes in the mapping between valency complementations and surface syntactic positions, the valency complementation expressed in a less prominent surface syntactic position is shifted into the more significant syntactic position (subject or direct object) of the second valency complementation. Whereas the prominent position is ‘multiplied’ either by syntactic means (coordination, as in (9b)), or by morphemic means (plural, as in (10b)), the less significant position is deleted from the resulted surface syntactic structure:

- (9) a. *Petr*.ACT-Speaker *svěřil* *Pavlovi*.ADDR-Recipient *své problémy*.PAT-Information – b. *Petr*.ACT,ADDR-Speaker,Recipient *a Pavel*.ACT,ADDR-Speaker,Recipient *si svěřili své problémy*.PAT-Information  
 Eng. a. Peter.ACT-Speaker revealed his problems.PAT-Information to Paul.ADDR-Recipient – b. Peter.ACT,ADDR-Speaker,Recipient and Paul.ACT,ADDR-Speaker,Recipient revealed their problems.PAT-Information (= one to another, reciprocally)
- (10) a. *Přítel*.ACT-Speaker *svěřoval své problémy*.PAT-Information *příteli*.ADDR-Recipient – b. *Přátelé*.ACT,ADDR-Speaker,Recipient *si svěřovali své problémy*.PAT-Information  
 Eng. a. The friend.ACT-Speaker revealed his problems.PAT-Information to his friend.ADDR-Recipient – b. The friends.ACT,ADDR-Speaker,Recipient revealed their problems.PAT-Information (= one to another, reciprocally)

We observe that in case of the verb *svěřit* ‘to reveal’ in (9a)-(9b) and (10a)-(10b), the situational content as well as the structural meaning are the same. The situational content is characterized by a set of the following participants: ‘Speaker’, ‘Recipient’ and ‘Information’. These participants are mapped onto the valency complementations ‘ACTor’, ‘ADDRessee’, and ‘PATient’ in the same way, respectively.

What differs these uses of the verb is the surface syntactic expression of ‘ADDRessee’: (i) it corresponds either to the syntactic position of indirect object (9a) and (10a), or (ii)

to the prominent subject position (9b) and (10b). In the later case, the change in the surface syntactic expression of ‘ADDRessee’ results in symmetric relation of ‘Speaker’ and ‘Recipient’, see Figure 2.



**Fig. 2.** The change in the mapping of the valency complementation ‘Addressee’ onto the surface syntactic positions of the verb ‘to reveal’ associated with syntactic diathesis.

Reciprocal constructions can be classified in various ways. E.g., they can be sorted according to which valency complementations are put in the relation of reciprocity. Let us introduce some frequent types:

- (11) **‘ACTor’–‘PATient’**  
*(Petr a Marie)<sub>rcp:ACT-PAT</sub> se líbali.*  
 Eng. (Peter and Mary)<sub>rcp:ACT-PAT</sub> kissed (each other).
- (12) **‘ACTor’–‘ADDRessee’**  
*Moji rodičové<sub>rcp:ACT-ADDR</sub> si dávali drahé dárky.*  
 Eng. My parents<sub>rcp:ACT-ADDR</sub> gave each other expensive gifts.
- (13) **‘ACTor’–‘ADDRessee’–‘PATient’**  
*(Oni)<sub>rcp:ACT-ADDR-PAT</sub> spolu nikdy o sobě otevřeně nehovořili.*  
 Eng. They<sub>rcp:ACT-ADDR-PAT</sub> have never talked with each other about themselves.
- (14) **‘ACTor’–‘DIRectional:to’**  
*(Oni)<sub>rcp:ACT-DIR3</sub> přistoupili k sobě.*  
 Eng. They<sub>rcp:ACT-DIR3</sub> have approached each other.
- (15) **‘ACTor’–‘ORIGin’**  
*(Oni)<sub>rcp:ACT-ORIG</sub> po sobě požadovali omluvu.*  
 Eng. They<sub>rcp:ACT-ORIG</sub> have asked apology from each other.
- (16) **‘ADDRessee’–‘PATient’**  
*Petr seznámil (rodiče a svou snoubenku)<sub>rcp:ADDR-PAT</sub>.*  
 Eng. Petr has introduced (his parents and his fiancée)<sub>rcp:ADDR-PAT</sub>.

**The representation of syntactic diathesis.** As in the case of grammatical diatheses, the changes in valency frame of verbs associated with syntactic diathesis concern only morphemic forms of the valency complementations. Similarly, these changes are regular enough to be described by syntactic rules which are stored in the grammar component of the lexicon. In the data component, only valency frames corresponding to the unmarked members of syntactic diathesis are listed. Then for each relevant lexical unit, the list of valency complementations which can be put in the symmetric relation are given in a special attribute.

For instance, the lexical unit of the verb *svěřit* ‘to reveal’ is represented only by the valency frame corresponding to the unmarked use. Then ‘ACTor’ and ‘ADDRessee’ which can be used reciprocally are listed in the special attribute which is ascribed to this lexical unit in the data component of the lexicon. In the grammar component, the following syntactic rule describing the changes in the surface syntactic expression of ‘ADDRessee’ complementation is formulated:

Rec.r	Unmarked	Marked
valency frame	ADDR <sub>dat</sub>	ADDR <sub>nom</sub>

**Table 2.** Rec.r rule for the reciprocity of ‘ACTor’ and ‘ADDRessee’.

For the representation of other types of reciprocity, similar syntactic rules are formulated by means of which valency frames for marked members of syntactic diathesis can be derived.<sup>3</sup>

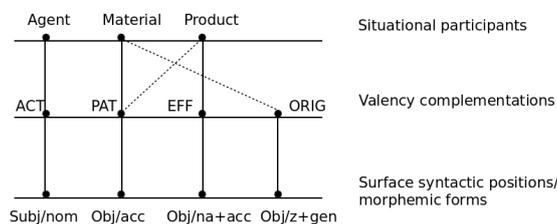
## 5 Semantic diatheses

Semantic diatheses and their representation are widely debated in [7] and [6]. Let us briefly recapitulate their characteristics. Semantic diatheses represent relations between two lexical units of a verb which are characterized by closely related situational contents: these situational contents usually consist of the same set of situational participants. However, their relations implicate a slight semantic shift. Moreover, different structural meanings, i.e., different valency frames, characterize the lexical units involved in these relations. Let us demonstrate this case on the pair of examples (17a)-(17b):

- (17) a. *Mlynáři*.ACT-Agent *semleli zrna*.PAT-Material *na mouku*.EFF-Product – b. *Mlynáři*.ACT-Agent *semleli ze zrna*.ORIG-Material *mouku*.PAT-Product  
 Eng. a. The millers.ACT-Agent ground wheat.PAT-Material into flour.EFF-Product – b. The millers.ACT-Agent ground flour.PAT-Product out of wheat.ORIG-Material

Both situational contents of the lexical units of the verb *semlít* ‘to grind’ consist of three situational participants, which show the same semantic properties and which are related by the same relations. However, we can observe that these relations result in different consequences: only example (17a), not (17b), implies holistic effect of the situational participant ‘Material’, i.e., the interpretation that the millers ground all wheat, see esp. [1] and [4]. This slight semantic shift is manifested by the change in the mapping of the situational participants onto the valency complementations, see Figure 3. The differences in the linking results in a different surface syntactic expression of the involved situational participants.

<sup>3</sup> A detailed description of changes in surface realization of individual complementations related to reciprocity can be found in [17].



**Fig. 3.** The changes in the mapping of the situational participants and the valency complementations of the verb ‘to grind’ associated with semantic diathesis.

Let us introduce frequent types of Czech semantic diatheses, see esp. [2]:

- (18) **Material-Product diathesis**  
 a. *Marta nakrájela chléb.*PAT-Material *na pět silných krajíců.*EFF-Product – b. *Marta nakrájela z chleba.*ORIG-Material *pět silných krajíců.*PAT-Product  
 Eng. a. Martha cut the bread.PAT-Material into five thick slabs.EFF-Product – b. Martha cut five thick slabs.PAT-Product from the bread.ORIG-Material
- (19) **Locatum-Location diathesis**  
 a. *Sedláci naložili seno.*PAT-Locatum *na vůz.*DIR3-Location – b. *Sedláci naložili vůz.*PAT-Location *senem.*EFF-Locatum  
 Eng. a. The farmers.ACT-Agent loaded hay.PAT-Locatum onto the truck.DIR3-Location – b. The farmers.ACT-Agent loaded the truck.PAT-Location with hay.EFF-Locatum
- (20) **Source-Substance diathesis**  
 a. *Slunce.*ACT-Source *vyzařuje teplo.*PAT-Substance – b. *Teplo.*ACT-Substance *vyzařuje ze Slunce.*DIR1-Source  
 Eng. a. The sun.ACT-Source radiates heat.PAT-Substance – b. Heat.ACT-Substance radiates from the sun.DIR1-Source
- (21) **Bearer-Location diathesis**  
 a. *V chrámu.*LOC-Location *zněl sborový zpěv.*ACT-Bearer – b. *Chrám.*ACT-Location *zněl sborovým zpěvem.*PAT-Bearer  
 Eng. a. Choral singing.ACT-Bearer was sounding in the church.LOC-Location – b. The church.ACT-Location was sounding with choral singing.PAT-Bearer

**The representation of semantic diatheses.** In contrast to grammatical diatheses and syntactic diathesis, the members of semantic diatheses correspond to separate lexical units. It implies that lexical units related by a particular type of semantic diathesis are represented by separate valency frames in the data component of the lexicon. These lexical units are interlinked by a relevant type of semantic diathesis.<sup>4</sup> Then in the grammar component, we formulate lexical rules indicating the mapping between situational participants and valency complementations.

Let us illustrate these principles on examples (17a)-(17b). In the data component of the lexicon, the lexical unit of the verb *semlít* ‘to grind’ in (17a) is represented by the valency frame (A) whereas the valency frame (B) is ascribed to the lexical unit in (17b):

<sup>4</sup> In [7] and [6], we proposed an adequate lexical-semantic representation of situational content of lexical units in the relation of semantic diathesis inspired esp. by [15]. However, the description of this representation goes beyond the scope of this paper.

- (A)  $ACT_{nom}^{obl}$   $PAT_{acc}^{obl}$   $EFF_{na+acc}^{opt}$   
 (B)  $ACT_{nom}^{obl}$   $PAT_{acc}^{obl}$   $ORIG_{z+gen}^{opt}$

Then in the grammar component, the lexical rule describing the changes in the mapping of the situational participants ‘Material’ and ‘Product’ onto the valency complementations is given, see Table 3.

Sem.r	Valency frame (A)	Valency frame (B)
‘Material’	PAT	ORIG
‘Product’	EFF	PAT

**Table 3.** Sem.r for the Material-Product diathesis.

Similar lexical rules can be formulated for other types of Czech semantic diatheses, see esp. [7] and [6].

## 6 Combination of diatheses

In our classification, diatheses of the same type cannot be combined together.<sup>5</sup> However, on certain conditions, different types of diatheses, i.e. grammatical, syntactic and semantic diatheses, can be combined. For instance, in case that a particular lexical unit related with another lexical unit by a certain type of semantic diathesis fulfils morphosyntactic and semantic conditions of applying a specific grammatical meaning, this lexical unit can create the marked member of a relevant type of grammatical diathesis.

Let us turn back to the lexical units of the verb *semlít* ‘to grind’ related by the Material-Product semantic diathesis, see examples (17a)-(17b). The grammatical meaning captured by the value of the grammateme *diagram* can be separately applied to these lexical units. Whereas the grammatical meanings represented by the value of the verbal grammateme *pass*, *deagent*, *rez1* and *rez2* are applicable to both perfective and imperfective lexical units (see examples (22)-(25)), the grammatical meaning *disp* is available only for the imperfective counterpart *semílat* ‘to grind’ (26a)-(26b). The value of the verbal grammateme *recip* is not applicable as it is conditioned by the occurrence of the situational participant ‘Recipient’ in the situational content of verbs.

- (22) a. (*Mlynáři*) bylo zrna semleto<sub>pass</sub> na mouku. – b. (*Mlynáři*) byla ze zrna semleta<sub>pass</sub> mouka.  
 Eng. a. Wheat was ground<sub>pass</sub> into flour (by the millers). – b. Flour was ground<sub>pass</sub> out of wheat (by the millers).  
 (23) a. Zrno se semlelo<sub>deagent</sub> na mouku. – b. Ze zrna se semlela<sub>deagent</sub> mouka.  
 Eng. a. ‘Wheat – *reft* – ground<sub>deagent</sub> – into flour.’ – b. ‘Out of wheat – *reft* – ground<sub>deagent</sub> – flour.’  
 (24) a. Zrno (již) bylo semleto<sub>rez1</sub> na mouku. – b. Ze zrna (již) byla semleta<sub>rez1</sub> mouka.  
 Eng. a. Wheat has (already) been ground<sub>rez1</sub> into flour. – b. Flour has (already) been ground<sub>rez1</sub> out of flour.

<sup>5</sup> In contrast to [3], we do not suppose that recipient passive and deagent grammatical diatheses can be combined together. Such combination is not supported by the corpus evidence

- (25) a. *Zrno (již) máme semleto<sub>rez2</sub> na mouku.* – b. *Ze zrna (již) máme semletu<sub>rez2</sub> mouku.*  
 Eng. a. Wheat has (already) been ground<sub>rez2</sub> into flour. – b. Flour has (already) been ground<sub>rez2</sub> out of flour.
- (26) a. *(Mlynářům) se zrno dobře semílalo<sub>disp</sub> na mouku.* – b. *Ze zrna se (mlynářům) dobře semílala<sub>disp</sub> mouka.*  
 Eng. a. ‘(For millers) – refl – wheat – well – ground<sub>disp</sub> – into flour.’ – b. ‘Out of wheat – refl – (for millers) – well – ground<sub>disp</sub> – flour.’

The main principles of the representation of grammatical and semantic diatheses (described in Section 3 and Section 5) remain the same: in the data component of the lexicon, two separate lexical units of the verb *semlít* ‘to grind’ are represented by the valency frames corresponding to the unmarked uses, see the valency frames (A) and (B) in Section 5. Then the applicability of the listed grammatical meanings is ascribed to these lexical units separately. On the basis of the syntactic rules, stored in the grammar component, the valency valency frames describing the marked uses of the units are derived.

Furthermore, if the valency frames corresponding to these units contain valency complementations which can be symmetrically used, lexical units can be used in reciprocal constructions. Thus semantic and syntactic diatheses can be combined together as well. E.g., in case of the lexical units of the verb *semlít* ‘to grind’, ‘ACTor’ and ‘BENefactor’ can be put in the relation of reciprocity:

- (27) a. *Mlynáři.ACT mu.BEN semleli zrno na mouku.* – b. *(Oni)<sub>rcp:ACT-BEN</sub> si (navzájem) semleli zrno na mouku.*  
 Eng. a. The millers.ACT ground wheat into flour for him.BEN – b. They<sub>rcp:ACT-BEN</sub> ground wheat into flour (for each other).
- (28) a. *Mlynáři.ACT mu.BEN semleli ze zrna mouku.* – b. *(Oni)<sub>rcp:ACT-BEN</sub> si semleli (navzájem) ze zrna mouku.*  
 Eng. a. The millers.ACT ground flour out of wheat for him.BEN – b. They<sub>rcp:ACT-BEN</sub> ground flour out of wheat (for each other).

As to the combination of grammatical and syntactic diatheses, such combination is available only in cases that a lexical unit of a verb exhibits the semantic and morpho-syntactic properties satisfying conditions of deriving both (i) a marked construction of a certain type of grammatical diathesis and (ii) a reciprocal construction.

Let us demonstrate these cases on the verbs *konzultovat* ‘to consult’ and *vyjednávat* ‘to negotiate’. Both these verbs allow to use reciprocally ‘ACTor’ and ‘ADDRessee’ (29b) and (30b), respectively. Moreover, the grammatical meaning described by the value *pass* and *disp*, respectively, of the verbal grammateme *diatgram* can be applied to these verbs (29c) and (30c), respectively. Then both linguistic means can be combined together (29d) and (30d), respectively.

- (29) a. *Jednotlivé kroky jedna strana.ACT konzultovala<sub>act</sub> vždy se stranou druhou.ADDR*  
 Eng. a. The one side.ACT consulted<sub>act</sub> individual steps with the other side.ADDR  
 b. *Obě strany<sub>rcp:ACT-ADDR</sub> jednotlivé kroky vzájemně konzultovaly<sub>act</sub>.*  
 Eng. b. Both sides<sub>rcp:ACT-ADDR</sub> consulted<sub>act</sub> individual steps together.  
 c. *Jednotlivé kroky byly s druhou stranou.ADDR konzultovány<sub>pass</sub> (stranou první).ACT*  
 Eng. c. Individual steps were consulted<sub>pass</sub> with the other side.ADDR (by the one side).ACT  
 d. *Jednotlivé kroky byly (oběma stranami)<sub>rcp:ACT-ADDR</sub> vzájemně konzultovány<sub>pass</sub>.*  
 Eng. d. Individual steps were consulted<sub>pass</sub> (by both sides<sub>rcp:ACT-ADDR</sub>).

- (30) a. *Já.ACT jsem s Janem.ADDR vyjednával<sub>act</sub>.*  
 Eng. a. I.ACT negotiated<sub>act</sub> with John.ADDR  
 b. *My<sub>recp:ACT-ADDR</sub> jsme (spolu) vyjednávali<sub>act</sub>.*  
 Eng. b. We<sub>recp:ACT-ADDR</sub> negotiated<sub>act</sub> (with each other).  
 c. *Dobře se mi.ACT s Janem.ADDR vyjednávalo<sub>disp</sub>.*  
 Eng. c. 'Well – refl – me.ACT – with John.ADDR – negotiated<sub>disp</sub>.'  
 d. *Dobře se nám<sub>recp:ACT-ADDR</sub> (spolu) vyjednávalo<sub>disp</sub>.*  
 Eng. 'Well – refl – us<sub>recp:ACT-ADDR</sub> – (together) – negotiated<sub>disp</sub>'.

However, in some cases, although the verbs satisfy conditions of both grammatical diatheses and reciprocity, the combination of these linguistic means is precluded. Let us exemplify this on the uses of the verb *slíbit* 'to promise' in examples (31a)-(31b), which are in the relation of recipient passive grammatical diathesis:

- (31) a. *Jan.ACT slíbil<sub>act</sub> Pavlovi.ADDR dárek.*  
 Eng. a. John.ACT promised<sub>act</sub> Paul.ADDR a gift.  
 b. *Jan.ADDR dostal od Pavla.ACT slíben<sub>recip</sub> dárek.*  
 Eng. b. 'John.ADDR – gave – from Paul.ACT – promised<sub>recip</sub> – a gift.'  
 b. *Pavel.ADDR dostal od Jana.ACT slíben<sub>recip</sub> dárek.*  
 Eng. b. 'Paul.ADDR – gave – from Jan.ACT – promised<sub>recip</sub> – a gift.'  
 c. *\*(Jan a Pavel)<sub>recp:ACT-ADDR</sub> dostali od sebe (navzájem) slíben<sub>recip</sub> dárek.*  
 Eng. c. \*'(John and Paul)<sub>recp:ACT-ADDR</sub> – gave – from themselves – (from each other) – promised<sub>recip</sub> – a gift.'

Although the valency complementations 'ACTor' and 'ADDRessee' meet the condition of semantic homogeneity, they cannot be used reciprocally (31c). In case of the marked construction of recipient passive grammatical diathesis, 'Agent' (corresponding to 'ACTor') is shifted from the subject position and this position is filled by 'ADDRessee'. However, in case of reciprocally used valency complementations, the valency complementation expressed in a less prominent surface syntactic position ('ACTor' expressed in the adverbial in (31b)) would be shifted to the syntactic position of the valency complementation which occupies a more significant position ('ADDRessee' expressed in the subject in (31b)). Thus to put 'ACTor' and 'ADDRessee' in reciprocity in (31b) would necessarily lead to the reshift of 'ACTor' into the subject. Such reshift would result in an ungrammatical construction (31c).

It follows that in the grammar component, it is necessary to determine the sequence of the rules which can be applied to individual lexical units so that the derivation of grammatically incorrect constructions would be prevented (see also [17]).

## 7 Conclusion

We have proposed a method of the representation of three types of Czech diatheses in the valency lexicon of Czech verbs, VALLEX. We have demonstrated that whereas grammatical and syntactic diatheses can be captured by syntactic rules, semantic diatheses can be represented by lexical rules. Finally, we show that on certain conditions the different types of diatheses can be combined together and that such combinations do not require any modifications of the proposed representation. The only necessary enhancement consists in a precise determination of the sequence of the syntactic rules. As to the future work, we intend to further examine the combinations of different types of diatheses.

## References

- [1] Anderson, S. R. (1971). On the role of deep structure in semantic interpretation. *Foundations of Language*, (7):387–396.
- [2] Daneš, F. (1985). *Věta a text: studie ze syntaxe současné češtiny*. Academia, Praha.
- [3] Daneš, F., Grepl, M., and Hlavsa, Z., editors (1987). *Mluvnice češtiny 3*. Academia, Praha.
- [4] Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, 67:547–619.
- [5] Kettnerová, V. and Lopatková, M. (2009). Changes in valency structure of verbs: Grammar vs. lexicon. In Levická, J. and Garabík, R., editors, *Slovko 2009, NLP, Corpus Linguistics, Corpus Based Grammar Research*, pages 198–210, Bratislava, Slovakia. Slovenská akadémia vied.
- [6] Kettnerová, V. and Lopatková, M. (2010a). Representation of Changes in Valency Structure of Verbs in the Valency Lexicon of Czech Verbs. In Bertinetto, P. M. et al., editor, *Proceedings of Verb 2010, Interdisciplinary Workshop on Verbs, The Identification and Representation of Verb Features*, pages 154–159, Italy, Pisa.
- [7] Kettnerová, V. and Lopatková, M. (2010b). The Representation of Diatheses in the Valency Lexicon of Czech Verbs. In Loftsson, H. et al., editor, *Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010)*, pages 185–196, Berlin / Heidelberg. Springer.
- [8] Kováčová, K. (2005). *Konverzívnost jako systémový vztah*. PhD thesis, Charles University in Prague, Prague.
- [9] Lopatková, M., Žabokrtský, Z., and Kettnerová, V. (2008). *Valenční slovník českých sloves*. Nakladatelství Karolinum, Praha.
- [10] Mel'čuk, I. A. (2004). Actants in Syntax and Semantics. *Language*, 12:1–66.
- [11] Panevová, J. (1999). Česká reciproční zájmena a slovesná valence. *Slovo a slovesnost*, 90:269–275.
- [12] Panevová, J. (2007). Znovu o reciprocitě. *Slovo a slovesnost*, 68:91–100.
- [13] Panevová, J. and Ševčíková, M. (2010). Annotation of morphological meanings of verbs revisited. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 1491–1498, Valletta, Malta. European Language Resources Association.
- [14] Panevová, J. et. al (manuscript). *Syntax současné češtiny (na základě anotovaného korpusu)*. Nakladatelství Karolinum, Praha.
- [15] Rappaport Hovav, M. and Levin, B. C. (1998). Building verb meanings. In Butt, M. and Geuder, W., editors, *The Projection of Arguments. Lexical and Compositional Factors*, pages 98–134. CSLI Publications, Stanford.
- [16] Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Reidel, Dordrecht.
- [17] Uřešová, Z. (manuscript). *Valence sloves v Pražském závislostním korpusu*. PhD thesis, Charles University in Prague, Prague.

# Translating Prepositions from Czech into Russian: Challenges for the Machine Translation

Natalia Klyueva<sup>1</sup> and Naděžda Runštuková<sup>2</sup>

<sup>1</sup>Institute of Formal and Applied Linguistics, Charles University of Prague

<sup>2</sup>Institute of Theoretical and Computational Linguistics, Charles University of Prague

**Abstract.** In this article we explore differences in preposition usage in Czech and Russian languages and how the Machine Translation (MT) system between the languages deals with prepositions. We focus on the errors that occur in preposition phrases. Our study involves research on a parallel corpus for theoretical evidence and analysis of the output of the rule-based Machine Translation (RBMT) system for closely related languages Česílko and the Statistical MT (SMT) system Joshua from Czech into Russian.

## 1 Introduction

This work presents the comparative analysis of prepositions and errors in preposition usage produced by the Machine Translation systems. We classify the mistakes, in some cases provide an explanation why the mistake occur. The study demonstrates an on-going work focused on analysis, so the system improvement itself lies out the scope of this paper and is planned to be made in future research.

Prepositions have been studied a lot both from the linguistic point of view and recently a few works on prepositions in the area of Natural Language Processing have appeared.

The theoretical basis of our work borrows several concepts from [1] and [8]. Importance of investigating prepositions is recognized also by computational linguists as the number of papers in this area grows rapidly. Just to mention the two papers on a very similar research: the proposal of how to handle prepositions in Machine Translation from English to Bengali is described in [7]; the paper [9] discusses the specific problem of translating locative prepositional phrases.

To the best of our knowledge no comparative study on prepositions in a computational sense for Slavic languages exist.

The paper is structured as follows. The Section 2 introduces theoretical introspection into the problem of preposition handling. Section 3 briefly presents the MT systems between Czech and Russian that our work will be applied at. Section 4 shows the most frequent types of errors connected with prepositions and Section 5 outlines a possible way for improving the translation of prepositions. Finally, we conclude in Section 6.

## 2 Prepositions in Czech and Russian languages: correspondences and differences

The role of the category of prepositions in Czech and Russian languages may be defined very similarly. The prepositions establish relationships between other words in a sentence. Most frequently they act in the sentence as a part of verbal valency and require certain case of the following noun or its substitute. Syntactically we can divide all prepositions to one-word prepositions and multi-word prepositions. By the origin, the prepositions may be classified into original and derived. One-word original Czech and Russian prepositions are in the majority of cases identical on the surface form as they have the same Slavic roots (eg. *na* – *на*, *s* – *с*, *z* – *из*). Moreover, the vocalization of Czech and Russian prepositions basically is very similar impact (*s/se* – *с/со*, *z/ze* – *из/изо*, *v/ve* – *в/во*, *k/ke/ku* – *к/ко*, more in [6]). In spite of the surface similarity the valency of prepositions can be different. For instance, a Czech original preposition *před* can govern nouns in Accusative and Instrumental, whereas the identical Russian *перед* only those in Instrumental. A preposition *proti* in Czech requires Dative case, and Russian *против* has a Genitive valency. We consider it to be one of the possible sources of the mistakes during the process of translation.

Formal similarity of prepositions does not automatically mean that they will be always translated with the corresponding ones. For example, in different contexts Czech preposition *na* with the main Russian equivalent *на* may be translated as:

1. **в** (*políbit na čelo* – *поцеловать в лоб*, *pracovat na univerzitě* – *работать в университете*)
2. **от** (*záviset na rodičích* – *зависеть от родителей*, *zemřít na otravu krve* – *умереть от заражения крови*)
3. **по** (*odborník na informační technologie* – *специалист по информационным технологиям*)
4. **для** (*lesk na rty* – *блеск для губ*)

Moreover, the students of Czech and Russian may be confused by a case of interlingual homonymy, such as *pro* (*for*) – *про* (*about*).

A lot of one-word prepositions are the part of the multi-word prepositions and they are translated in a certain way as a part of whole these lexical units: *na rozdíl od* – *в отличие от*, *ve vztahu k* – *по отношению к*.

czech preposition	total frequency	main equivalent	the frequency of equivalent	percentage
v/ve <i>L/A</i>	40915	в/во	26671	65
do <i>G</i>	7178	в/во	3599	50
s/se <i>I/G</i>	11810	с/со	5426	46
pro <i>A</i>	7247	для	2957	41
na <i>A/L</i>	24373	на	8303	34
po <i>L/A</i>	3770	после	1164	31
z/ze <i>G</i>	11237	из/изо	3103	28
k/ke/ku <i>D</i>	11280	к/ко	3015	27
za <i>A/I/G</i>	6290	за	1502	24
o/vo <i>L/A</i>	9623	о/об/обо	2059	21

**Table 1.** Probability of a proper translation of prepositions

On the materials of the Czech-Russian parallel tagged corpus UMC 0.1<sup>1</sup> we have measured the probability with which 10 most frequent Czech prepositions are translated as the respective frequent Russian variants. The values of equivalent translation varied from 65 per cent (e.g. *v* into *в*) to 21 per cent (*o* into *о*). For more results see Table 1.

The usage of the other kind of translation equivalent is generally caused by phraseological reasons (the case of multi-word prepositions and prepositional phrases as *v létě – летом*, *konvice na čaj – чайник*), or the lexical choice of verb with the certain valency (*pracovat na něčem – работать над чем-то*).

In several cases the prepositional phrase may be translated as non-prepositional. This fact might be supported by statistics: the total number of prepositions in Russian in this corpus is 184920 whereas on the Czech side there are 148931 prepositions.

So it is evident, that simple substitution of a Czech preposition with its Russian equivalent will in many cases result in a mistake. Next we will describe the MT systems between Czech and Russian and see the percentage of errors made with respect to preposition usage.

### 3 Machine translation between Czech and Russian

Machine Translation belongs to one of the most popular areas of Natural Language Processing. Nowadays lots of rule-based systems for different languages exist. Tools for statistical Machine Translation (e.g. Moses) allow to make a MT system for each language pair having a parallel corpus. For Czech and Russian languages both types of systems were created, and in our work we decided to explore both the rule-based system which is simpler to find and explain the source of the errors and to compare the results with statistical MT. We should note, that Statistical MT systems handle a prepositional phrase correctly if it was seen in the training data in an appropriate context. For the SMT we are not able to predict if the mistake in preposition usage will

<sup>1</sup> <https://ufal.mff.cuni.cz/umc/cer/>

occur or not, and it is more easy to do this for the rule-based system. For example, we do not know how an SMT will translate *lesk na rty* (it depends if the phrase was seen in the training data), and Česílko system is supposed to make a mistake: *блеск на губы*.

### 3.1 Česílko – a MT for Slavic Languages

Česílko [2] is a simple RBMT system that was first developed for translating from Czech into Slovak, later for more distant languages – Polish, Lithuanian and Russian were introduced into the system. As Czech and Slovak are very closely-related languages the system achieved a good translation quality. The MT systems created for other more distant language pairs scored significantly worse, moreover, they were only experimental, so the dictionary was limited and they could not be tested on a randomly chosen text. It became evident, that a simple word-for-word translation architecture will not be enough for those pairs, and the transfer module on a shallow syntax level need to be created. Several modules were added to the system – among them those responsible for morphological analysis of Czech sentences and synthesis of Russian, a shallow syntactic transfer module.

First experiments for Czech-to-Russian MT are described in [3]. The system demonstrated rather a low translation quality mainly because it used a rather low-quality dictionary automatically extracted from a parallel corpus. Now as the dictionary is cleaned up, we have decided to make some other improvements of the system based on the types of mistakes that were found after the translation of a test set. The latter presents 1000 Czech sentences in the newspaper genre that have reference translation in several languages, among which there is Russian. 100 sentences out of a test set were manually flagged with error markers. It turned out that a special marker for a preposition should be used, because there were lots of mistakes connected with preposition usage. These mistakes are described in detail in Section 4.

### 3.2. Joshua

Joshua [5] is an open-source platform for making experiments in a Parsing-Based Statistical Machine Translation. This toolkit contains several modules that process parallel texts in two languages and outputs the translation model enriched with syntactic analysis. We used the parallel corpus UMC mentioned above to train the model. Then we explored the output of Joshua in the comparison with Česílko results.

## 4 Analysis of the MT output with respect to prepositions

In the process of error flagging we have found out that almost 50% (Česílko) and 36% (Joshua) of all evaluated sentences have mistakes related to the preposition usage. In this section we provide a short analysis of these errors.

We have divided all the errors into several groups: preposition ambiguity, verbal valency, errors in cases of noun phrases after prepositions and phraseology errors. Further on we give examples of those errors for both MT systems.

### Preposition ambiguity

This type of errors is connected with the ambiguity of prepositions which causes that they can be translated in many different ways into the other language, including the possibility of non-prepositional variant of translation. For that reason the one-to-one translation fails and the statistical tool shows the better result than Česílko. On the other hand, the SMT is a black box resulting in unpredictable output, so it can either solve this problem properly or to make mistakes.

In the example sentence (1) the Czech preposition *po* 'to' was translated according to the dictionary, which has only one most frequent variant of translation (*po* – *после*) 'after' in Česílko not considering the second meaning of the preposition (*po* – *до*) 'to'. Example (2) demonstrates the same type of the mistake made by Joshua, where the preposition *в* instead of *для* should be used.

Česílko: (1) *od brutálního bití novinářů po nezákonné zadržování* – *от жестокого избияния журналистов ERR:после незаконного задержания* (lit. 'from brutal beatings of journalists to illegal confinement')

Joshua: (2) *obnovit svou důvěryhodnost pro izrael* – *восстановить свою репутацию ERR:для израиля* (lit. 'to restore its credibility with Israel')

### Errors concerning the verbal valency

Usage of a wrong preposition in many cases originates in a difference in valency frames. A verb is translated properly, still the valency frame in Czech and Russian has some differences. The borderline between the errors caused by the valency frames and by the ambiguity interpretation sometimes may be rather vague.

Česílko: (3) *hlasovaly pro ústavní změny* – *проголосовали ERR:для конституционные изменения* (should be translated as *за*; lit. 'voted for constitutional changes')

Joshua: (4) *k mírovým vyhlídkám v jižní asii jen přispěje* – *к мирному перспективы в южной азии только укрепит* (no preposition needed, NP should be in Accusative case; lit. 'will only help improve peace prospects in South Asia')

In these two examples both MT systems did not processed properly the valency frames of verbs 'to vote' and 'to strengthen' which are different in Czech and Russian.

### Errors in cases of nouns governed by prepositions

Another type of errors occurs in a morphological form of a noun phrase governed by a preposition. For instance, in Czech the preposition *proti* 'against' requires a noun in Dative whereas its Russian analogue *против* governs Genitive case. That difference caused an error demonstrated in (5). The similar error which is even more frequent occurred in translation by Joshua. It was not only copying wrongly the corresponding cases from Russian word forms, but also illogically put the words in other case. In the example (6) the preposition 'in' requires a noun phrase in Genitive case in Czech and Accusative in Russian, Joshua assigned Genitive case to the noun and adjectives accept for the last word of the NP that was unexpectedly in Instrumental.

Česílko: (5) *proti náboženským.DAT radikálům.DAT – против ERR:религиозным ERR:радикалам* (case should be genitive, not dative); lit. 'against religious radicals')

Joshua: (6) *lákat mladé lidi do svých.GEN vzdělávacích.GEN a sociálních.GEN sítí.GEN – привлечь молодых людей в своих.GEN образовательных.GEN и социальных.GEN сетью.INS* (all words after the preposition should be in Accusative; lit. 'lure young people into their educational and welfare networks')

### Errors in phraseology

The usage of phraseological units requires a certain predicted translation, so they can not always be translated literally. Otherwise, the construction can sound awkward or have another meaning, e.g.

Česílko: (7) *stát za svým – стоять за своим* (lit. 'to stand for one's own')

Joshua: (8) *boj kdo s koho – борьба кто с кого* (lit. 'fight who with whom')

Table 2 demonstrates the rough statistics of the preposition error rate.

Type of an error	#of errors in Joshua	#of errors in Česílko
Preposition ambiguity	17	26
Errors in verbal valency	6	10
Errors in a case of a governed noun	38	16
Errors in phraseology	5	5
Total number of prepositions in the set	185	

**Table 2.** Errors in preposition usage in the Czech-Russian MT system calculated on 100 test sentences

From the Table 2 we can see that the two most significant differences between the types of errors made by the systems lay in the area of preposition ambiguity and the case of a governed noun. Česílko made more errors while translating prepositions with their first most frequent equivalent. Joshua did not use the cases of governed nouns properly, because of the data sparseness that is especially high for the morphologically rich languages. Joshua could not generate the proper case forms just because they were not seen in the training data.

These weak points require improvement in order to increase the score of proper translated prepositions and prepositional phrases.

## 5 Suggestions for the system improvement

Regarding the mistakes described in the previous section we propose a measure aiming at the improvement of the preposition translation via our MT system.

The additional information on phraseology and verbal valency is planned to be introduced into the dictionary as this problem is very specific for a concrete word. Verbal valency for Czech and Russian is addressed in [4], the list of differences presented is supposed to handle those most frequent cases of preposition incorrespondence. The list of Czech and Russian multi-word prepositions is available in [6]. As for the preposition disambiguation and the right choice of a case after a preposition we plan to introduce a set of rules for preposition transfer for the most frequent of them. The rules will be rather primitive in a sense that they do not use a semantic notion of time, location or some other meaning of a prepositional phrase.

It is obvious, that the rule-based approach will not cover all the cases. So for the rule-based system more linguistic data is needed, whereas the SMT system will be improved if more and more parallel data is added.

## 6 Conclusion

In this article we have described an initial phase of both theoretical and practical studies of prepositions in Czech and Russian. We have showed several types of differences between Czech and Russian prepositional constructions and examined errors in preposition usage that occur in automatically translated texts. Although the languages of our interest are closely-related, the number of mistakes connected with prepositions is relatively high. So we have outlined the possible direction in creating an improvement scheme for the MT systems that will be realized in future. Additionally we made a comparison of rule-based and statistical systems with respect to preposition translation. It turned out that generally SMT scored better when translating prepositions, and there were several types of preposition errors that were typical for the RBMT and SMT.

## 7 Acknowledgements

This work was supported by the grant MSM 0021620838. We would also like to thank Dan Zeman for his help with Joshua.

## References

- [1] Čermák, F. (1996). Systém, funkce, forma a sémantika českých předložek. In: Slovo a slovesnost, 57, pp. 30-46.
- [2] Hajič, J., Kuboň, V., Hric, J. (2000). Česílko – an MT system for closely related languages. In *ACL2000, Tutorial Abstracts and Demonstration Notes*, pp. 7-8. ACL, Washington.
- [3] Homola P., Klyueva N., Bojar O. (2009). Towards a Rule-Based Machine Translation System Between Czech and Russian. In: *Formal Description of Slavic Languages*, University of Potsdam, Potsdam, Germany, pp. 37-38.
- [4] Klyueva, N., Kuboň, V. (2010). Verbal Valency in the MT Between Related Languages. In: *Proceedings of Verb 2010, Interdisciplinary Workshop on Verbs, The Identification and Representation of Verb Features*. Pisa, Italy, pp. 160-164.

- [5] Li, Z., Callison-Burch, Ch., Dyer, Ch., Ganitkevitch, J., Khudanpur, J., Schwartz, L., Thornton, W., Weese, J., Zaidan, O. (2009). Joshua: An Open Source Toolkit for Parsing-based Machine Translation. In Proceedings of the Workshop on Statistical Machine Translation (WMT09).
- [6] Nadykta, N. (2007). The Use of Prepositions in Czech and Russian According to Parallel Corpus Data. Master Thesis.
- [7] Naskar, S. K., Bandyopadhyay, S. (2006). Handling of Prepositions in English to Bengali Machine Translation. In the Proceedings of the Third ACL-SIGSEM Workshop on Prepositions; April 3, 2006, Trento, Italy.
- [8] Šiganova, G. A. (2002). Sistema leksičeskich i frazeologičeskich predlogov v sovremennom russkom jazyke. Dissertation thesis.
- [9] Trujillo, A. (1992). Locations in the machine translation of prepositional phrases. In: Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-92); Montreal, Canada. Laval (Quebec): CWARC; 1992:13-20.

# Extracting Verbs with PP/NP Variation from the Large 3-gram Corpus\*

Mikhail Kopotev<sup>1</sup>, Natalia Kochetkova<sup>2</sup> and  
Eduard Klyshinsky<sup>3</sup>

<sup>1</sup>University of Helsinki

<sup>2</sup>Moscow State Institute of Electronics and Mathematics

<sup>3</sup>Keldysh Institute of Applied Mathematics RAS

**Abstract.** In any language with a case system, certain verbs govern variation in noun/prepositional phrases (Cf. Eng. *to write him / to him*). In Russian, with its rich morphology, this variative government is so widespread as to be traceable back to both the diachronic process and semantic shifts in the modern language. To launch a deep linguistic analysis of this phenomenon, one should first identify verbs with variative PPs/NPs. In this article, we present a corpus-based approach to automatically extract such verbs from a large corpus. The algorithm is based on searching 3-grams such as ‘V + (Prep) + N’ and comparing them to each other with the help of some filters. The results achieved for variative non-prepositional NPs appear to be too noisy, however, and require further development. Those for variative PPs, on the other hand, are much more accurate and can serve in linguistic research.

## 1 The objectives

In Russian, verbal government varies, as in *стрелять по уткам / уток / в уток* (lit. *to shoot at ducks / ducks / into ducks* ‘to shoot at ducks’). Investigations of Russian syntax indicate that the government of specific verbs or verb classes may eventually change [1-4]. Based on these studies, three types of variation can be established. Below are examples from 18th to 20th-century texts.

- Variation in noun phrases (hereinafter NP): *знать грамоту / грамоте, прощать автора / автору, бросить камень / камнем*
- Variation in noun and prepositional phrases (hereinafter PP): *поздравлять камер-юнкером / с камер-юнкером, смеяться вам / над вами*
- Variation in prepositional phrases: *намекать на опасность / об опасности, рассказывать о битве / про битву.*

Some verbs may form mixed variations in which an NP competes with more than two PPs (Cf. *надеяться выздоровления / на выздоровление / о выздоровлении / в выздоровлении*). In all cases, a verb governs an NP or a PP, so it is more accurate to speak of different surface syntactic groups rather than merely of case variation. One question that arises is how to mark the boundaries between variative NPs/PPs and

---

\* This work has received partial support from RFBR grant #10-01-00800.

groups with similar though not identical meanings. The meanings of competitive forms, for example, quite clearly differ in pairs (1) and (2) below. Cf.

1. *стрелять уток / по уткам* lit. *to shoot ducks / at ducks*
2. *платить адвокату / за адвоката* lit. *to pay for a lawyer / to a lawyer.*

In example (1), the meanings of the NP and PP are almost identical in that both describe the same situation with slight differences in lighting the scene. The pair (2) obviously describes two different scenes with two different subjects. To distinguish such cases, we relied on the following rough criterion: the groups are considered variative if they cannot function simultaneously in the same sentence with the same nouns. According to this criterion, it is impossible to accept the clause *\*Стреляю уток по уткам* (lit. *\*I shoot ducks at ducks*); although the clause *платю адвокату за адвоката* (lit. *I pay a lawyer for a lawyer*) has at least one sensible interpretation, rather rare in real life though.

One more question also comes to our attention here, and that is the question of language norms. Language mavens consider *приехать с Москвы* (lit. *to come out Moscow*) instead of *приехать из Москвы* (lit. *to come from Moscow*), for example, to be utterly unacceptable. However, many (if not most) Russians use the preposition *с* without hesitation. By avoiding the question of purism, one can more impartially forecast the future development of the language.

As far as we know, existing methods for extracting such verbs from texts are incomplete and unreliable. This article proposes an algorithm for automatically identifying variability based on the 3-gram corpus, which was built up from a large array of textual data [5]. The working hypothesis is that a large corpus is well suited to identifying a sufficient number of verb-preposition-noun chains in which a certain verb functions with different forms of the same noun, whether an NP or a PP. The following example clarifies this point. Our algorithm is considerably more complicated than a simple search for accusative or dative forms of *утка* ‘duck’ after the verb *стрелять* ‘to shoot’. Rather, we identify chains containing different forms of the same lemmas following a certain verb (e.g. *стрелять [уток / по уткам] [зайцев / по зайцам]* ‘to shoot’ [at ducks / ducks] [at rabbits / rabbits]).

The large array of textual data, which we used to compile the 3-gram corpus, contains about 759 million running words in texts of various genres, thus provides enough data for this kind of work (Table 1 presents the data in more detail).

Sources	Volume, million words
M. Moshkov’s library	680
Lenta.ru	39
PCWeek	28
Membrana.ru	12
<b>Total:</b>	<b>759</b>

Table 1.

To annotate the data with morphological tags, we used the Russian tagger Crosslator, created at the Keldysh Institute of Applied Mathematics, Moscow. The 3-grams have been extracted according to the following algorithm. Input texts are processed and patterns are identified in every sentence where:

**Noun group** = {Preposition (b.)} + {Ordinal numeral (p.)} + {Possessive pronoun (p.)} + {Adverb (p.)} + {Adjective (b.)} + {coordinate conjunctive (p.)} + {Adjective (b.)} + Noun (b.)

**Verbal group** = {Noun group (b. if in initial position)} + {Adverb (b.)} + Verb / Participle / Gerund (b.) + {Adverb (b.)} + Noun group (b.)

The ‘b’ indicates that the given word appears in the 3-grams corpus, whereas ‘p’ indicates that the word does not. ‘{ }’ indicates optional availability. “Noun + Verb” combinations are marked in the output file in order to facilitate searches for the inverted word order. In the event the algorithm meets an unexpected or morphologically ambiguous word, analysis of this combination ceases. As a result, the n-grams obtained are stored in two indices:

- The index of n-grams in their original forms, as they appear in the texts
- The index in which, for each combination of V + N, Ger + N, Part + N, V + Adv, Gen + Adv, Part + Adv, and N + Adj, the running words are standardized into lemmas, and its frequency is calculated.

To extract verbs with case variation, we used the second index.

## 2 Algorithms and filters

### 2.1 The basic algorithm

Below is a description of the algorithm for the automatic extraction of verbal 3-grams with variative government. This is an implementation of the corpus-based approach [6], which claims that all retrieved data must, in theory, be relevant, though the quality of the algorithm remains to be evaluated in terms of recall and precision (see Part 3). The algorithm is as follows:

1. Search for a verb  $v_1$
2. Take its lemma  $V_1$
3. Search for a noun  $n_1$  and preposition  $prep_1$  (if any)
4. Take the given noun  $n_1$  and preposition  $prep_1$  and their lemmas  $N_1$  and  $PREP_1$
5. Search for a second verb, whose lemma ( $V_2$ ) is equal to  $V_1$  ( $V_2 = V_1$ )
6. Verify whether  $V_1=V_2$  and  $N_1=N_2$  and ( $(n_1 \langle \rangle n_2$  and  $PREP_1=PREP_2$ ) or ( $n_1=n_2$  and  $PREP_1 \langle \rangle PREP_2$ )
7. Save n-grams “ $V_1 + (prep_1) + n_1$ ” and “ $V_1 + (prep_2) + n_2$ ” in the index.

In its initial stage, we selected only n-grams with a direct word order, so the noun-verb order thus far remains unanalyzed. All n-grams were completed with lemmas of nouns and verbs; the frequency of each combination was also indicated in the index.

## 2.2 Filters

The initial algorithm yields a fairly large percentage of noise, so we complemented it with a series of filters in an effort to improve the output.

- In the algorithm, the regular alternation of singular and plural forms was considered as two different forms. To eliminate this error, we added control over number and case; if two nouns belong to the same lemma and have the same case, but differ in number, then the n-gram under processing is excluded.
- Verbal forms with *-ся* endings always cost too much in Russian NLP, insofar as the automatic disambiguation of passive and reflexive forms is impossible. What is more difficult in our case is that ambiguous *-ся* forms often govern different NPs and PPs (Cf. *причитать по мужу / по муже* ‘to lament for / over the husband’, but *причитаться мужу* ‘to be due to the husband’). As a result, these *-ся* forms were completely excluded from the search function at this stage.
- One more filter is needed for multiword expressions (MWEs), because they often include syntactic words, which are homonymous to nouns (Cf., e.g., *в течении* (‘in the stream’, PP) vs. *в течение* (‘during’, MWE)). For this reason, they have been collected and excluded from the final index as well.
- Yet any abbreviations written in capital letters were also omitted, which reduced the noise dependent on homonyms such as *МОК* (Eng. I[nternational] O[lympic] C[ommittee]) and *мок* ‘soaked’.

The results obtained were organized according to verb lemmas, so that the algorithm produces chains containing verbs with variation in governed phrases.

## 3 Evaluation

### 3.1 Results

Applying the filters substantially improved the results. Further development continued in classifying them into smaller groups to which more specific filters could also be applied. These groups contain:

1. Variations of NPs (*бросить камень / камнем*)
2. Variations of NPs and PPs (*исчислить в деньгах / деньгами*)
3. Variations of different prepositions in PPs (*приехать с Москвы / из Москвы*)
4. Variations of the same prepositions in PPs (*набрать в поиск / в поиске*).

The results for the first and second groups were definitely skewed. The main problems were related to the homonymy of nominative and accusative cases, and competition between accusative and genitive cases in interrogative clauses. Below is a typical example of erroneous processing:

НАЙТИ;КНИГА;;книга (e.g. *Покупателя найдет книга*)

НАЙТИ;КНИГА;в;книге (e.g. *Нашел в книге*)

НАЙТИ;КНИГА;;книгу (e.g. *Нашел книгу*)

НАЙТИ;КНИГА;;книги (e.g. *Не найди книги*)

Both problems seem unsolvable using the given corpus. One possible solution could be based on the wider context, which enables one to resolve case ambiguity as well as to control negation in the clause. The simplest way to accomplish this is to assign the nearest ambiguous nominative/accusative form, which is left to the verb, as nominative. Unmarked word order in Russian renders the correct assignment of the case highly predictable. This is the task to be solved, although the results for groups 3 and 4 are much more likely :

ПЕРЕЙТИ;ИСПОЛЬЗОВАНИЕ;к;использованию;

ПЕРЕЙТИ;ИСПОЛЬЗОВАНИЕ;на;использование;

ВСТАТЬ;РЕЙД;на;рейд;

ВСТАТЬ;РЕЙД;на;рейде;

Moving ahead, we decided to improve the results by adding more filters. The main filter is based on the frequency of constructions in which variative phrases compete. This approach is based on the Construction Grammar theory [7, 8]. The overall logic of this algorithm is as follows: In given pairs 1-2 and 3-4,

(1) ВСТАТЬ;РЕЙД;на;рейд;

(2) ВСТАТЬ;РЕЙД;на;рейде;

(3) ВСТАТЬ;ЯКОРЬ;на;якорь;

(4) ВСТАТЬ;ЯКОРЬ;на;якорь;

The 1-2 pair is formally identical to the 3-4 pair. That is the construction ‘V + Preposition + Noun (Acc/Prep)’. Both pairs contain the same verb *встать* ‘embark’ and the same variative PPs “*на* ‘on’ + Acc. case” and “*на* ‘on’ + Prep. case”. The algorithm looks for occurrences of similar constructions in the 3-gram corpus, and if a) 3-grams occur where verbal and noun lemmas are equal in pairs, and b) a preposition in a pair coincides with a preposition in another pair, and c) the cases of the noun in the first pair coincides with the cases in the second, then the chains are considered correct. Additionally, we used a frequency threshold cut-off so as to require at least three iterations for each coincidence in order to reduce accidental coincidences.

### 3.1 Validation

Using the 3-gram corpus, we found it impossible to obtain more or less acceptable results for non-prepositional NPs. The evaluation shows that the percentage of erroneous output is higher than 70%. As noted above, any further development should take into account a wider context in order to control the case after negation, as well as

the nominative/accusative ambiguity. These results have not been validated in detail due to their obvious weakness. For variative PPs, the results contain substantially fewer errors. To evaluate this output, we verified it with the following two standard parameters: recall and precision.

**Precision** is the proportion of documents retrieved that are relevant to the results obtained and was calculated for one hundred randomly selected verbs in each group. For those with different prepositions in PPs, variative government was confirmed in 77%; noise constituted 23%. In the group with the same prepositions in PPs, the relevant results were confirmed in 82%; noise constituted 18%.

More important in this respect a **recall**, or the fraction of results that are relevant to all relevant results. Obviously, precision errors can be removed from the list manually, while missing verbs cannot, by any means, be added to it. A recall is calculated using a “gold standard”; a list compiled from sources [9-10], which includes 102 variative verbs; the recall rate is 91.8%, which is considered acceptable.

All in all, the biggest part of the corpus, the M. Moshkov’s library, provides 3 309 verbs which govern different prepositions in PPs and 1174 verbs which share the same prepositions in PPs. With the correction for resultant noise, the list yields 2 548 and 962 verbs, respectively. If one compares these results to those in the existing researches [1-4, 9-10], their number exceeds that of previously published results by a factor of ten. This provides a sound starting point for further linguistic research.

## References

- [1] Maier, I. (1997). *Verbalrektion in den “Vesti-Kuranty” (1600–1660). Eine historisch-philologische Untersuchung zur mittelrussischen Syntax*. Uppsala.
- [2] Maier, I. (2006). *Verbalrektion in den “Vesti-Kuranty” (1600–1660). Teil 2: Die präpositionale Rektion*. Uppsala.
- [3] Dubrovina, L. (2002). *Variativnoe glagol’noe upravlenie v russkom yazyke pervoi treti XIX veka*, Uppsala.
- [4] Ferm, L.: (2005) *Variativnoe bespredlozhnoe glagol’noe upravlenie v russkom yazyke XVIII veka*. Södertörn-Huddinge.
- [5] Klyshinsky, E.S., Kochetkova, N.A., Litvinov, M.I., Maksimov, V.Yu. (2010). Avtomaticheskoe formirovanie bazy sochetaemosti slov na osnove ochen’ bol’shogo korpusa tekstov.  
<http://www.dialog-21.ru/dialog2010/materials/pdf/29.pdf>
- [6] Mukherjee, J. (2004). The state of the art in corpus linguistics: three book-length perspectives. In *English Language and Linguistics*, 8(1), 103-119.
- [7] Fried, M. & Östman J.-O. (eds.). (2004). *Construction Grammar in a cross-language perspective*. Amsterdam: John Benjamins.
- [8] Goldberg, A. (2006). *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.
- [9] Rozental’, D.E. (1986). *Upravlenie v russkom yazyke*. Moscow: Kniga.
- [10] Shvedova, N. Yu. (ed.). (1980), *Russkaya grammatika*. Moscow: Nauka.

# **On the Question of Homonymy and Polysemy in the Lexicographical Practice of the Russian Language Semantic System in its Development Modelling**

Irina M. Nekipelova

Izhevsk State Technical University

**Abstract.** The article is devoted to the electronic model of a historical-etymological dictionary. In our research we suggest a new perspective on differentiation and correlation of the semantic word-formation and derivation notions based on the language history. This differentiation promotes a more clear understanding of the homonymy and polysemy phenomena and makes it possible to identify and systematize the processes of the semantic changes in the history of the language. Lexical-semantic word representations as the units of lexicographic description are based on these processes.

## **1 Introduction**

Research of the Russian semantic system changes is an important element in the reconstruction of the historical development of the language. And it is the semantics of linguistic units, lexicalization and phraseologization of certain ideas that express the mental content of the system. Unfortunately, it is impossible to trace the development of the semantic system using existing historical and etymological dictionaries, though they contain invaluable lingo-historical and etymological data. Printed dictionary model presupposes the static nature of data presentation, i.e. from the lexicographical units (lexemes in absolute use or phraseologized unit) to the contexts of their usage. The developing electronic historical-etymological dictionary model allows making an attempt to represent a word in its historical development, because it is built on a different principle, i.e. from the context to the lexicographic unit (lexeme in the absolute use or phraseologized unit).

## **2 Problems and principles of compiling a dictionary's model**

### **2.1 Material of the dictionary**

This dictionary project presupposes the lexicographic representation of the Old Russian language units, since this linguistic system was used as a means of legal, economic and business writing for the state. Also being used nationwide, it became the basis for the literary form of Russian – the highest form of a national language. Though initially translated from Greek and later own religious writing played a considerable role in the Old Russian language development, it was not the only factor that determined its progress.

Today the basic data for lexical- and grammatical- semantic description are the lexical units of XI-XIV century business written texts. One of the earliest samples of business written records are records of scribes, artists and bookbinders, which were collected and prepared for printing by L.V.Stolyarova [Свод 2000], Smolensk documents of the XIII-XIV centuries, prepared for printing by R.I. Avanesov and published by the Academy of Sciences of the USSR in 1963 [Смоленск 1963] and XI-XV centuries ancient Russian princes' ustavs (most of them are available in later copies), prepared for printing by Ya.N. Shchapov and published by the Academy of Sciences the Institute of History of the USSR in 1976 [Древнерусские 1976]. These samples of ancient literature underlie the database which contains the transcriptions of the ancient manuscripts. To our regret, today's electronic written texts corpus does not contain any of the above mentioned important government documents, but they could be very helpful in the reconstruction of the early Russian written language. The importance of this cannot be underestimated since we can only suppose the processes that were characteristic of the preliterate period of the Russian language, but can't assert them.

In the future we are planning to update the dictionary database with business manuscripts and written chronicles of the XV-XVII centuries. The texts of the manuscripts should be presented in integration with the Slavic culture. It should be noted that many linguists erroneously believe that the Old Russian language is a dead language, and was not used for formal writing. This is not true, because the Old Russian language is the form in which the Russian language existed at the early stages of its development, and at that period the writing system of the language emerged. The Old Russian language period was followed by the Middle Russian language, which was used in the Middle Ages (XV-XVII centuries). It existed until at the beginning of the XVIII century, when the literary form of the language appeared. That's why we don't include in our dictionary database any of the written records of the XVIII century.

## **2.2 Semantic model of the dictionary**

The theoretical framework for the creation of the electronic-semantic lexicographic model was produced by the analysis of the samples of business [Nekipelova 2005] and religious [Nekipelova 2010] written records. The linguistic research of the ancient and medieval written samples makes it possible to assert that the semantic system of the Russian language in its history progressed in a vector-oriented way from the concrete word, directly naming denotation, through the formation of abstract definitions, which name notions abstracted from agent, to notions denoting abstract concepts [Nekipelova 2005]. The research of the semantic system of the Old Russian language presupposes the study of its two important components: the lexical and grammatical semantics. Historical semantics is an integral part of the semantics on the whole, which allows enriching the well-known linguistic picture of the Slavic world known to modern scientists, which is reflected in the texts of the ancient literature through differences in word functioning. As Florensky metaphorically said, the word is really a way of the energy concentration of a semantic concept [Florensky V. 2. 1990: 292]. Describing the semantics of a word and its links and relations to other words, we first of all rely on the context in which the word is used, and also on vari-

ous linguistic dictionaries data, which fix the usage of the researched words in similar or different contexts.

The process of the creation of the historical and etymological dictionary – is the process of lexical and grammatical semantics modeling, which is presented in the diachronic aspect, because a dictionary is a model “a priori”. There are several factors that add to the complexity of dictionary modeling. First of all, the main problem lies in the construction of the semantic model, because it inevitably must take into account the data from all levels of the language, while semantics itself is not the language level. Semantics is the content of the language. Secondly, when modeling the semantics, one must take into consideration the process of continuous language change, the change of the semantic content of which is the most important variable element. So the model must combine both temporal and spatial relations, vertical (diachronic) and horizontal (synchronic) correlations as well as the connections of the linguistic units within the system (the model). Thirdly, it is worth mentioning that the word in the present denotes something only in a certain context, which hampers fixing all possible word meanings and its relations to other linguistic units, because the texts that survived till our times may not reflect all of the word’s meanings and relations, which were realized in one or another period of the language development. Fourth, when we interpret words relationships used in ancient texts, we cannot be sure of absolute reliability of this analysis, because nowadays we consider the words semantics through the prism of the semantic relations among words that exist in the modern language, which might not exist in the earlier stages of the language development. Fifth, any modeling including semantic one uses abstraction, idealization and formalization, which eliminates or minimizes deviations from the general change pattern, while variations and offshoots are the essence of the semantics development.

Reflecting some relevant and important from the viewpoint of the research properties of the original and omitting unimportant viewpoints, the model plays the role of an abstract idealized object. Any model is based on the hypothesis about the original’s possible structure and is the functional analogue of the original. That allows transferring knowledge from the model to the original. Ideally, the model must be formal (i.e., initial objects and relations which are linked to them must be explicitly and unambiguously defined) and have explanatory power (i.e., it must predict the unknown before, but possible behavior of the original). The same is applied to the semantic model. Semantics runs through all levels of the language. That is why when building semantic models one should take into consideration different processes happening at various language levels. This is the fundamental difference between semantic models and all other linguistic models. Here the model’s fullness is not the main factor, because fullness here is not related to universality. While one of the main criteria of any, including linguistic, model is the prediction of the unknown but possible behavior, which should be later confirmed by observation or experiment. The proposed in this article semantics modeling represents the development of typical semantic description structure, while the fullness of this structure is conditioned by the individual characteristics and relations of words.

### 2.3 The semantic description of lexicographic units: homonymy and polysemy

It should be noted that in the process of the electronic presentation of lexicographic units one of the biggest challenges both on the theoretical and practical levels is detecting historical and semantic words relations in the language and text. That is connected with the ambiguity of the terms *polysemy* and *homonymy* correlations. However, the linguistic data analysis indicates that these two terms do not oppose each other, because they do not denote the same phenomena, but denote various objects and phenomena of the language, which have not obtained adequate substantiation and distinction in linguistics, since it was not taken into account that *it is a sign that is polysemantic, not a word*. As Kolesov formulated it, when we talk about the word polysemy, we mix the notions of word and sign, of the part and the whole. [Kolesov, 2002, pp. 19].

For more complete understanding of the polysemy and homonymy phenomena in the language history it is necessary to distinguish the concepts *semantic word-formation* and *semantic derivation*. Even though the term *word-formation* is quite easy to understand, the term *derivation* might cause some difficulties. *Semantic derivation* is considered in linguistics as a special case of *derivation* on the whole. However, the *derivation* is generally understood as *deviation* from the normal or baseline condition, *change* of the direction, *branching*. So observing the general scientific trend and the tradition in the term *derivation* defining, linguistic term *semantic derivation* should be used to *describe the semantic processes that lead to deviations from the original meaning of the word and its changing*.

Reflecting on the above mentioned we come to the conclusion that the term *semantic derivation* is not opposed to the term *semantic word-formation*. Even though in certain areas of semantic words relations these two terms can intersect, it is not correct to equal these concepts. Both semantic derivation and semantic word-formation reflect diachronic processes in the language and that is why we used them to describe historical events. However, each term should be used to do research on some definite linguistic processes.

Let's give the definition of each term. *Semantic word-formation* is the process characteristic of any language system that leads to the appearance of a new word, which is called a *semantic derivative* and also leads to the formation and development of the *homonymy* phenomenon. When a *sign* is polysemantic (first of all its graphic structure) it is associated with two or more words with different denotations which were derived from the original word. Through the *semantic derivation* process a word starts possessing some additional *semantically derived values*, "*by-meanings*", *semantic connotations*, i.e., semantic derivation may be called as the *process of semantic capacity of a word expanding*. It leads to the appearance of *semantic syncretism* and further to the *process of the semantic syncretism* disintegration that in its turn leads to appearance of the so-called *polysemy* phenomenon. According to the derivation *directions* different types of derivation may be distinguished: *metonymical and metaphorical changes of the semantic structure of a word*.

Semantic derivation and semantic word-formation phenomena are closely related. They reflect different stages of the word's semantic development and carry out different functions in the semantic systems construction. *Semantic derivation* is a quantitative factor in semantic changes. It means that the appearance of a word's new

connotations increasingly expands its semantic capacity (though in some exceptional cases we should speak not about the expanding of a semantic capacity, but of narrowing). The semantic derivation process first of all happens in the speech rather than language, because the actualization of certain connotations is due to the context in which the linguistic unit is used. Thus, the context plays an important role in the functioning and disintegrating of semantically-filled elements (syncretisms). *Semantic word-formation* is the result of a qualitative leap, which leads to appearance of new linguistic unit with its own value that is fixed in the language. In this case context doesn't determine the actualization of one or another connotation of a used linguistic unit, but choice of a word from a number of units that possess similar properties both in meaning and form. Thus, *semantic derivations* are the semantic *changes* in the semantic structure of words. *Semantic word-formation* is *development* of the semantic system of a language within the bounds of which a new element as a linguistic unit starts functioning. As a result, systemic relations between the producing and derivative elements that appear are *types of metonymisation* and *models of metaphorisation*. More new concepts of differentiating value are accumulated, and as Kolesov said systemic relations that regulate the processes of semantic derivation come into existence [Kolesov, 2002, p. 96]. Thus, semantic syncretism is the process of meanings *expanding*; semantic derivation is the process of meanings *changing* and semantic word-formation is the process of meanings *distinction*. However, it is not always possible to distinguish the transition of derivation to word-formation, because it cannot be always predicted whether the semantic derivation will result in the semantic word-formation, i.e. will lead to the appearance of a new linguistic unit or meaning expanding; or the changing of the semantic structure of the word will have only contextual actualization.

So, the word in its semantic changing and the language system in its semantic development undergo the following stages: 1) originally the word exists and functions in language and speech in its original objective meaning; 2) further development of the semantic system and language on the whole leads to the forming of semantic syncretisms, i.e. linguistic units with the permanently expanding semantic capacity of meaning as a result of semantic derivation process, i.e., changes in the semantic structure of a word; 3) later in the XI-XIV centuries as the result of the semantic derivation and the actualization of connotations in contexts process the disintegration of the semantic word syncretism of a word is actualized; 4) from the end of XIV to the beginning of XV centuries the process of semantic word-formation actualized, which finalized the phenomenon of the semantic syncretism disintegration with the appearance of new lexical units. These units are semantic derivatives that entail the development of the semantic system of the language. It should be noted that not all connotation have already reached this stage.

The language system always develops, but in order to do research on semantics it is important to fix intermediate steps that are static in synchrony. But dialectic processes with respect to language are universal. Static must be considered as an element of dynamic, as well as synchronicity sections should be interpreted as diachronic stages. So, all the semantic changes must be regarded as the stages in semantic system development (the diachronic approach), leading to the temporary static position of the semantic system (the synchronic approach).

### 3 Conclusion

Thus, one of the challenges of electronic historical-etymological dictionary modeling is the presentation of semantic changes in the context of the linguistic system development. The dictionary system, which is oriented to presenting lexicographical units not as isolated elements, but as functioning in the text in integrity with other units, both the notions of polysemy and homonymy in the language and text should be reflected. So, for describing word meanings in XI-XIV centuries' texts, it is necessary to provide the integrated representation of word's connotations that represent semantic syncretism. To represent word meanings in the later times texts it is important to reflect the actualization of one or another connotation in the certain context. To determine word meanings used in texts of XV-XVII centuries it is necessary to show the value of derivatives, which have developed in the semantic system of the language.

Such representation of language elements and word values is a very modern problem. The proposed representation is the most accurate one and fully reflects the processes of the semantic changing in semantic structure of a word and also allows seeing the direction of the development of the semantic system in making a word autonomous, i.e. from a text unit to the unit of the language. The main goal in creating the electronic historical-etymological dictionary lies in the structural and systemic representation of the semantic system on its way towards the literary form of the Russian language.

### References

- [1] Древнерусские княжеские уставы XI-XV вв. М.: Наука, 1976. – 240 с.
- [2] Колесов, В. В. (2002). Философия русского слова. СПб.: ЮНА.
- [3] Некipelова, И. М. (2005). Метонимическая и метафорическая деривация в истории русского языка (на материале памятников деловой письменности XI-XVII веков). Автореферат... канд. филол. наук. – Казань: КГУ.
- [4] Некipelова, И. М. (2010). К вопросу о развитии грамматической и лексической семантики в истории языка: синкретичное слово в контексте. Слово и текст в культурном сознании эпохи: Сборник научных трудов. Часть 4. Вологда: Легия, с. 221-227.
- [5] Смоленские грамоты XIII-XIV веков. Под ред. члена-корреспондента АН СССР Р. И. Аванесова. (1963). М.: Изд-во Академии наук СССР.
- [6] Столярова, Л. В. (2000). Свод записей писцов, художников и переплетчиков древнерусских пергаментных кодексов XI-XIV вв. М.: Наука.
- [7] Флоренский, П. А. (1990). Столп и утверждение истины. М.: Правда, Т. 2.

# Historical and Etymological Electronic Dictionary System Modelling

Irina Nekipelova and Elvira Zarifullina

Izhevsk State Technical University

**Abstract.** The article describes the concept of electronic historical-etymological dictionary compiling. It considers the nature of the data and the composition of texts included to the database, the structure of the system and its individual components: the visualization module, the input module and the queries module.

## 1 Introduction

Knowledge accumulation that presents the world picture has been one of the top areas of human activity for a long time. The mankind history evidenced the construction and then destruction of huge libraries, which confirms the special importance of accumulating knowledge in books for humanity. Among the publications intended to preserve the already existing experience and knowledge dictionaries undoubtedly take the first place, because dictionaries present the accumulated information in the processed and systematized form. The language system constantly changes, as well as the picture of the world is being modified all the time. That means that the content of dictionaries should be replenished and improved. The idea of how to create dictionaries is constantly improved and new theories of lexical units' representation and description are suggested. However, the data that the existing dictionaries contain is of great value for modern science. Using and analyzing the existing dictionaries data, as well as combining data from both historical and modern dictionaries may favor the systematization of the information about the worldviews of the ancient Slavs and their language system. In our research we plan to analyze the available information and add to it the peculiar features word usage in the ancient written samples of different genre affiliation and linguistic composition. This idea is fundamental for the creation of the historical-etymological dictionary, lexical units of which will be the data form written samples of the XI–XVII centuries. There are lots of written records of the ancient literature, which will form the basis of the database. The lexical units of the dictionary database will be the linguistic units of the Old Russian language, the language on the basis of which the Russian literary language developed. That is why as the material for the research we have selected primarily business written samples, and particularly written samples of the XI–XVII centuries. A number of existing samples of business records, particularly of the XVI–XVII centuries are enormous. They vary in genre: legal codes, diplomatic (letters, contracts), administrative (department records, cadasters and census books), judicial (investigative cases, lawsuits, petitions and judicial decisions), economic (inventories, receipt and expenses books) and other materials. In most cases they represent important and valuable linguistic sources, because they show in more detail than book texts the peculiar features of the spoken

language of those times. One of the earliest samples of business written records (records of scribes, artists and bookbinders) were collected and prepared for printing by L.V. Stolyarova [Svod 2000], XIII–XIV centuries Smolensk documents were prepared for printing by R.I. Avanesov and published by the Academy of Sciences of the USSR in 1963 [Smolensk, 1963], the ancient Russian princes' ustavs, XI–XV centuries (most of them have survived to nowadays later copies), prepared for printing by Ya.N. Shchapov and published by the Institute of History of the Academy of Sciences of the USSR in 1976 [Old Russian 1976]. These samples of the ancient literature underlie the database that contains the transcription of ancient manuscripts. In the future we are planning to add chronicle texts to the database.

## 2 Second section

Electronic historical and etymological dictionary is a computer system consisting of a database that contains the ancient written samples transcriptions, and specially structured entries; system describing the meanings of independent and phraseological (formulaic expressions) lexical units; the search module that allows selecting articles taking into consideration morphology; and the texts and lexicographical visualization module.

### 2.1 Database

The main part of the system is a specialized database LexBase, realized in MySQL and containing the texts of the XI–XIV centuries business literature samples. Initially this database includes the texts of the early business writing samples: clerks's records, artists and bookbinders records, which were collected and prepared for printing by L.V. Stolyarova [Svod 2000], Smolensk documents of XIII–XIV centuries [Smolensk 1963], as well as the ancient Russian princes' ustavs of XI–XV centuries. [Old Russian 1976]. In the future we are planning to enlarge the database by including other samples of business writing, including the XV–XVII century's samples.

The dictionary database will also contain the historical data of the following dictionaries with links to these dictionaries: Sreznevsky I.I. Materials for the Dictionary of the Old Russian Language, Dictionary of Russian XI–XVII centuries, Dictionary of Old Russian XI–XIV centuries; and etymological dictionaries: Fasmer M. Etymological Dictionary of the Russian Language, Etymological Dictionary of the Slavic languages; Proto-Slavic word collection (Eds. Trubachev O.N.), Chernyh P.Ya; Historical-etymological dictionary of the modern Russian language, Ivanov V.V., Shanskaya T.V., Shanskiy N.M. A short etymological dictionary of the Russian language.

Currently the database consists of the tables containing texts with their descriptions, lexemes their relations to other units of the language. Thus, the tables contain both the actual data and the results of the authors' research. One of the characteristic features of the database is the availability of the lexemes' semantic characteristics table. It is a flexible means for describing the semantic content of the word with the maximum degree of completeness that allows taking into account the peculiarities of the word in the texts of different genre and time period. The database is designed to allow further expanding of the "sub-bases" texts, which helps to avoid being limited

to groups of texts of some concrete time period, genre, or subjects, and also has the ability to store different views of the text. For example, the adequate representation of the original and the presentation of text divided into linguistic units, lexemes to make it more readable.

## 2.2 Requests

One of the main functions of any information retrieval system is the possibility to process user's requests and to produce the necessary result. Output data can be texts, dictionary entries, or snippets. Output data can be texts, dictionary entries, or snippets. Our project of the historical and etymological dictionary suggests the realization of three types of search requests:

*Metatext search* presupposes searching the database according to the user-specified characteristics (time of creation, a probable place of creation, author, genre, etc.). As a result, a list of relevant texts is formed – texts that match the query. Metatext features include the title, author (if known), exact or approximate dating of the written sample, the approximate or exact location of the sample, the genre of the text and some other features. Features can be entered fully or partially, by using, for example, only dating or genre. All the texts of samples included to the database dictionary will have a hypertext structure. In this structure, the information is not systematized in accordance with the bibliographical principle in which the unit of storage and retrieval is the text itself, but in accordance with the object graphical principle, which means that the stored unit is information containing additional information about the retrieval object. The text is the compound object, each element of which has a separate description. In this connection, a hypertext search allows to access a detailed description of the interested units of the relevant text with hyperlinks – parts of the hypertext document related to the relevant dictionary entries.

*Full-text search* is carried out throughout the relevant text when a user enters the lexemes of interest to the search string. The result of this query is a snippet. It is the visualization of the selected text, which contains interesting lexeme or verbal formula, entered to the search box.

If needed to expand the system functionality adaptive search that looks through all the documents of the system can be implemented.

## 2.3 Input data module

The database filing will be carried out automatically with the auxiliary module "AutoAdd", which was developed by the authors of the project. This module extracts lexemes from the marked text, and inserts them into the database, filtering out the ones similar to the already existing units of the database. The module assesses the degree of the coincidence of the lexical and phraseological unites by asking the linguist operator. If this units set of the relevant lexemes is already registered, it will create a corresponding link between the unit of the text and a lexeme available in the database. So the principle is achieved without redundancy of information. This mechanism helps the linguist to mark the text effectively, solving the problem of matching characters (homonyms and polysemous words with the actual value) in the text. It makes it possible also to ensure the establishment of the necessary connection

between lexemes in order to more quickly access synonymous units based on the coincidences of word meaning interpreting. Still the most common variant is when the word has a synonymous one in the system, but the system does not show the synonymous relationship between the lexemes. At this moment it is linguist-expert who separates and characterizes the semantic relations between the words. All the above mentioned procedures are applied to differentiate homonyms. In these cases it is impossible to talk about the automated differentiation of such complex phenomena as homonymy, polysemy and synonymy. But within the bounds of this module it is possible to realize in future the scheme of data input, based on the associative searching, able to identify accurately such “bottlenecks” in the semantic description of linguistic units (the dictionary objects).

#### **2.4 Visualization module**

Ergonomic interface and logical data visualization is one of the biggest tasks in system design. Visualization usually solves the problem of visualizing large amounts of heterogeneous data, since the system must quickly process large volumes of available information in the database. The focus of electronic dictionaries and systems working with large amounts of textual information is on the completeness of the interest data representation, despite the fact that it is impossible to provide all arrays of data on the screen at a time. The used visualization methods in the system can provide the user with the information in a compressed form as a result of the selection of the necessary data from its set. In our electronic historical and etymological dictionary visualization module is a control panel consisting of three windows: visualization window of the text adequate to the original, visualization window of the text divided into linguistic units, which are dictionary elements, and the windows of the dictionary entry. These windows can be displayed all at once, and in various quantitative ratios, depending on the choice of the user. The minimal representation is one box (any of the three windows, including the only window with the dictionary entry), the maximum performance – three windows in sequence: 1) The original text – 2) the text divided into lexical units – 3) the dictionary entry. In cases where the manuscript has already reflected the division into words, the representation of the text will only be in one window, since this type is adequate to the original. Such representation of texts and descriptions of the constituent elements of language makes the system a flexible tool for linguistic assessment and texts analysis. Thus, the visualization module of historical-etymological dictionary is a unique tool for studying the Old and Middle Russian language texts, because it has no analogues and allows the users to customize the screen in accordance with their interests.

### **3 Conclusion**

Historical and etymological dictionaries exist, but they are realized only in printed form. Often they are limited to the synchronic approach of the lexical units consideration [electronic; Varina 2004]. Moreover, the database of electronic dictionaries contains only one of the many lexeme characteristics (dictionaries, dictionaries of synonyms, paronymies, phrasebooks, etc.).

Dictionaries, accumulating a lot of knowledge about the word, have appeared recently and nowadays are actively developing (a multilingual dictionary «Lingvo», project Gramota.Ru). The realization of the electronic historical and etymological dictionary will create a powerful tool for linguistic analysis of not only the samples of XI–XVII centuries business writing, but of the Old Russian language in the diachronic aspect on the whole.

## References

- [1] Древнерусские княжеские уставы XI–XV вв. М.: Наука, 1976., 240 с.
- [2] Смоленские грамоты XIII–XIV веков. Под ред. члена-корреспондента АН СССР Р.И. Аванесова. (1963). М.: Изд-во Академии наук СССР.
- [3] *Столярова, Л.В.* (2000). Свод записей писцов, художников и переплетчиков древнерусских пергаментных кодексов XI–XIV вв. М.: Наука.
- [4] *Варина С.Н., Васильева О.В., Волков С.С.* (2004) Словарь обиходного русского языка Московской Руси XVI–XVII веков. Выпуск 1: А – Бязь. СПб.: Наука.
- [5] *Иванов В.В., Шанская Т.В., Шанский Н.М.* (1971) Краткий этимологический словарь русского языка. – М.
- [6] Словарь древнерусского языка (XI–XIV вв.) (1988-2004). В 10 т. /Р.И. Аванесов. – М.
- [7] Словарь русского языка XI–XIV XVII вв. (1975-2004.) Вып. 1-27. М.: Наука
- [8] *Срезневский И.И.* Материалы для словаря древнерусского языка, тт. I-III. Спб.,(1893-1903) / М., (2003).
- [9] *Фасмер М.* (1964-1973) Этимологический словарь русского языка. Пер. с нем. Т. 1-4. М.: Прогресс.
- [10] *Черных П.Я.* (1994) Историко-этимологический словарь современного русского языка. М.
- [11] Электронный словарь русского языка XVIII века / Фундаментальная электронная библиотека РУССКАЯ ЛИТЕРАТУРА И ФОЛЬКЛОР. Режим доступа:  
<http://feb-web.ru/feb/sl18/slov-abc/03/sl301804.htm>
- [12] Этимологический словарь славянских языков. (1974-1995 (изд. продолжается)) Праславянский лексический фонд /под ред. О.Н. Трубачёва. Т. 1-22.

# A Web-based Morphological Tagger for Bulgarian

Aleksandar Savkov<sup>1</sup>, Laska Laskova<sup>1</sup>, Petya Osenova<sup>1</sup>, Kiril Simov<sup>1</sup>  
and Stanislava Kancheva<sup>1</sup>

<sup>1</sup>Linguistic Modelling Department, IICT, Bulgarian Academy of Sciences, Bulgaria  
savkov@bultreebank.org, laska@bultreebank.org,  
petya@bultreebank.org, kivs@bultreebank.org,  
stanislava@bultreebank.org

**Abstract.** In this paper, we present a web-based morphosyntactic module for Bulgarian, which includes a statistical tagger and a lemmatizer. Both tools are implemented as a pipeline. It comprises an SVM-based tagger, a lexicon look-up component, a set of morphosyntactic context rules and a lemmatizer. The input and output of each component is defined according to the WebLicht format. Thus, a better compatibility and a better workflow is ensured with the NLP architectures for Bulgarian and other well-processed languages.

## 1 Introduction

Recently, the NLP community focused on two perspectives: integrating the existing resources and tools for various languages, and making them available on the web for the public community. The time of compiling various resources and tools is far from over, although there already exist a lot of such resources and tools. However, the existing ones are not accessible or integrated into usable application architectures. For that reason, the pan-European CLARIN initiative put as its main goal the ‘communication’ among all differing resources as well as their ‘applicability’ to the area of humanities.

To be in line with these most recent requirements, our group has started to integrate its resources in pipe-lines, and to equip them with the necessary web services. Our goal is not just to publish the available resources and tools, but also to improve them according to the experience we gained during their exploitation within several projects. We have started with a language infrastructure for Bulgarian, as developed within BulTreeBank project [9]. This infrastructure comprises several language resources and tools: a text archive of more than 100 mln. running words, a morphologically annotated corpus of 1 mln. tokens, a syntactic Treebank of 214 000 tokens, various lexicons, a morphological analyzer, partial grammars, named entity analyzers. In the span of several years all of them were developed further. In order to provide access to the latest developments, we decided to re-implement some of our tools. This re-implementation relies on new approaches, new parameters and new linguistic knowledge. We aim at providing at least the following web-based services accessing our resources and tools:

- Language Resources Services:
  - Concordance over plain text;
  - Concordance over annotated text.
- Language Technologies Services:
  - Morphological analysis of documents provided by the users;
  - Lemmatization of documents provided by the users;
  - Syntactic analysis of documents provided by the users.

In this paper we describe our language pipeline for the first two of the language technology services, which incorporates a statistical SVM-based tagger, a large morphological lexicon, a rule-based component for correction of the morphological annotation and a lemmatizer. The pipeline is implemented on the base of several language modules. For each of these modules we have implemented web services. In this way, they might be easily integrated also in other pipelines. In fact, we are developing alternative modules for some of the tasks in order to provide better flexibility opportunities to the user (such as, providing various statistical taggers within the pipeline). The presented pipeline is made available on the web as a free service, and also synchronized with the WebLicht format in order to be compatible with similar pipelines for other languages. The WebLicht is a web-based service environment that allows the users to integrate and use language resources and tools [4].

The paper is structured as follows: In Section 2 the state-of-the-art taggers for Bulgarian are presented and briefly discussed. Section 3 describes the architecture and the implementation details of the SVM-based tagger. Section 4 focuses on the error analysis at the various stages of processing. Section 5 concludes the paper.

## 2 State-of-the-art morphological analysis for Bulgarian

In this section we present the systems for morphological tagging of Bulgarian that we are aware of. The first attempts were just applications based on morphological lexicons. For example, morphological analysis has been performed by the morphological dictionary for Bulgarian [6]. However, in this approach the morphosyntactic ambiguities remained unresolved. After having compiled also some annotated gold standard texts with resolved ambiguities, the researchers have directed their efforts towards the creation of automatic taggers, which to handle the disambiguation task.

There exist several morphosyntactic taggers for Bulgarian. Our group has created or participated in the creation of some of them, such as [7] and [2], while there were also other attempts, such as [1]. In [7] a gold standard of 2500 sentences has been used for training a neural network system. These sentences were selected with the aim to demonstrate the most frequent ambiguities per sentence. A rule-based component was also added before the automatic analysis. The architecture of the system is hybrid, since it uses both components – the rule-based one and the neural network one. The neural network was trained to solve the hardest cases when the input con-

tains a lot of ambiguities. Thus, we expected it to perform better on a simple input. This is why in the hybrid system we first applied the rule-based component, which solved some of the ambiguity problems with 100% accuracy. Then the neural network was applied to solve the rest of the problematic cases. The accuracy for the part-of-speech feature only is 95.25%. When all the morphosyntactic characteristics were included in the evaluation, the accuracy dropped to 93.17%. The underlying tagset was large – more than 600 tags. In order to cope with the sparseness of the data, we applied two approaches – one was the selection of a corpus with a lot of ambiguities (see above), and the second was to encode the input to the neural network in the form of a vector of morphosyntactic features and to learn the co-occurrences among them. Another approach was performed in [2]. The tagset was simplified on the base of contributing local features. A smaller tagset was designed by reducing the features which do not contribute to the disambiguation task. The experiments have been performed with a reduced tagset – about 108 tags. The accuracy was improved to 94.43 % over all features.

[1] uses a rule-based method on a bigger training corpus. The implementation is in FSA. It achieves a precision of 98.4 %. Thus, all the mentioned taggers rely on the application of a rich morphological dictionary and linguistic rules. However, they differ in the automation method, and the training corpus.

Additionally, several machine learning POS taggers were trained on the BulTreeBank morphologically annotated corpus. Atanas Chanev trained TnT, SVM and Example-based taggers. The parameters files are available at our web site<sup>1</sup>. The Tree Tagger was trained by Julien Nioche within the European project LIRICS. The result is available from the TreeTagger site<sup>2</sup> and it is also included in the distribution of GATE system. None of these taggers reported the accuracy of the tagging.

In spite of the fact that these taggers are constructed as pipelines, they are standalone applications. Only some of them are available to a wider audience, but they are not equipped with a web service. Our own tagger from 2001 is outdated in at least two respects – the tagset was further developed and does not match the training data anymore; and, new training data is available.

Thus, we decided to implement a new version of the morphosyntactic tagger reusing a similar architecture – a hybrid approach, which includes a machine learning component and a rule-based component. Also, we have implemented a lemmatizer on the top of the tagger. The next section presents the architecture and the implementation of the components behind the services.

### 3 Implementation of the language pipeline

In this section we present the implementation of the pipeline performing the morphosyntactic tagging and the lemmatization of Bulgarian texts. From our own experience as well as from others' work, we directed our efforts towards a hybrid architecture. It

---

<sup>1</sup> <http://www.bultreebank.org/taggers/taggers.html>

<sup>2</sup> <http://www.ims.uni-stuttgart.de/projekte/complex/TreeTagger>

comprises a rule-based component and a machine learning component. The combination of these components and the definition of features to be learned provide the possibility of implementing several systems. For example, in our system, based on a neural network, the features to be learned were the morphosyntactic features of the word forms in a window, sliding over the test sentences. In this case, the input has been already annotated with all possible analyses for each word form on the base of the morphological lexicon. Thus, the application of the rule-based component was done prior to the application of the neural network. In other approach, the input to the machine learning component is the clean text without any annotation. In this case, the features include the words within the text, their endings, tags from the gold standard corpus, etc. However, the rule-based component is run after the annotation, produced by the machine learning component. Here our work complies with this second approach. In other words, the morphological lexicon and the disambiguation rules repair the output of the machine learning component. On the top of the result a lemmatization module has been implemented, which is based on rules.

### 3.1 Statistical PoS tagging

The machine learning component for POS tagging is built as a wrapper around a tool developed by Jesús Giménez and Lluís Màrquez called SVMTool – [3]. The SVMTool is a generator of sequential taggers based on Support Vector Machines (SVM). It is based on the SVM<sup>Light</sup> implementation of Vladimir Vapnik's Support Vector Machine [10] created by Thorsten Joachims [5]. Its central idea is to turn the problem of POS tagging into a binary problem and then to use the SVM<sup>Light</sup> tool for performing classification before making a decision. The binary representation of the problem is realized by splitting the problem of tagging a word with one of the tags from a tagset into multitude of problems consisting of deciding whether a word should be tagged with a specific tag or not. To achieve that, the SVMTool trains an SVM for each tag in the tagset and then uses them to determine the tag for the word in focus.

The flexibility of the SVMTool allows it to be trained for an arbitrary language as long as it is provided with annotated data. For the purpose of the training, a 650 000 token annotated corpus from the BulTreeBank manually annotated corpus was prepared in the training format – with a single token-tag pair on each line. The results we achieved with the optimal training configuration ranged from 89 % to 91 % accuracy. The minimal and maximal values within the above mentioned range depend on the genre of the text. On newspapers the values are closer to 91 %, while when tried on prose they are closer to 89 %. All these results were produced on the base of the standard parameters provided with SVMTool directly. We used the best trained model as a baseline for the Bulgarian tagger, and also as input for the rule-based component.

### 3.2 Lexicon look-up and morphosyntactic rules

The task of the next component of the morphosyntactic tagger is to correct some of the erroneous analyses made in the process of the statistical POS tagging. The typical errors, made by the SVM-based analysis, are described in Section 4 below. Since the statistical tagger does not use the information from the lexicons, the tag suggested by the SVM differs in some features from the real possible tags found in the lexicon. The correction of the wrong suggestions is performed by two sources of linguistic knowledge – the morphological lexicon and the set of context based rules. In the process of repairing we used as much as possible from the information provided by the SVM tagger. The context rules are designed in such a way that they aim at achieving higher precision even at the cost of low recall. The lexicon look-up is implemented as cascaded regular grammars within the CLaRK system – see [8]. The lexicon is an extended version of [6] and covers more than 110 000 lemmas. Additionally, a set of gazetteers were incorporated within the regular grammars. After the lexicon look-up application, 97.01 % of the tokens (in the test corpus) received morphological analysis from the lexicon and the gazetteers.

In the examples below a WebLicht encoding of the data is presented, extended with some additional attributes for the information from the lexicon. Thus, each token is represented by the element `<ns2:token>`; an `@aa` attribute with all possible morphosyntactic analyses is assigned to each token by the lexicon look-up; the analysis produced by the SVM Tool is encoded as an `@svm` attribute; and the final analysis is encoded as an `@ana` attribute. Some tokens do not have `@aa` attributes because they are not in the lexicon. In this case the attribute `@ana` has the same value as the `@svm` attribute. If the attribute `@aa` is attached to a token, we assume that the value of the `@ana` attribute has to be among the values of the `@aa` attribute. Thus we have the following clashes between the SVM-based component and the rule-based component:

1. The attribute `@aa` contains only one tag. In this case this tag is the value of `@ana` attribute;
2. The attribute `@aa` contains several tags and the value of `@svm` tag is among these tags. In this case, the context rules are applied. If they suggest a different tag from the one in attribute `@svm`, it is considered as the true one, and it is assigned as a value of attribute `@ana`;
3. The attribute `@aa` contains several tags and the value of the `@svm` tag is not among these tags. Again, context rules are applied. If they suggest a tag, then it is considered the true one. Otherwise, one of the tags in `@aa` is selected by intersection of the features encoded in them and the tag suggested by the SVMTool. In case this intersection is empty, we leave the ambiguous information for later processing, including the tag, suggested by the SVMTool.

About 30% of the errors are corrected simply by consulting the inflectional dictionary and the gazetteers. The context rules take into account the value of `@aa`

attribute and the intrasentential morphosyntactic context. The number of rules is 70. These rules are implemented as value restriction constraints in the CLaRK System. Each constraint has a target section and a source section. The target section defines the token for which the rule will take a decision, and its context. For example, the phrase 'големи котешки очи' (big cat eyes) is represented as:

```
<ns2:token aa="A-pi" ana="A-pi">големи</ns2:token>
<ns2:token aa="A-pi;Amsi" svm="Ncfpi">котешки</ns2:token>
<ns2:token aa="Ncnpi" ana="Ncnpi" svm="Ncnpi">очи</ns2:token>
```

In this case, the target of the rule is the element:

```
<ns2:token aa="A-pi;Amsi" svm="Ncfpi">котешки</ns2:token>
```

and the other two elements are the context which triggers the application of the agreement rule. In this case, the morphologically ambiguous form 'котешки' (cat's or cats'), which can be analyzed as a plural adjective (tag "A-pi") or as a masculine singular adjective (tag "Amsi"), is assigned the first value. The result is:

```
<ns2:token aa="A-pi" ana="A-pi">големи</ns2:token>
<ns2:token aa="A-pi;Amsi" ana="A-pi" svm="Ncfpi">котешки</ns2:token>
<ns2:token aa="Ncnpi" ana="Ncnpi" svm="Ncnpi">очи</ns2:token>
```

With respect to their accuracy, we consider the rules to be "sure" not in the sense that they achieve the highest possible recall, but in the sense that they achieve near 100 % precision when tested on the BulTreeBank corpus (over 1 000 000 tokens). Some rules do not have 100% precision, but we include them in the set if the precision is higher than 95%. The proportion between the rules with 100% precision and the others is 4:1 to the advantage of the first type. Here is one example of a rule of such a type. The word form 'че' (that) is ambiguous and despite the fact that it usually functions as a subordinate conjunction (Cs), it could be tagged also as an emphatic particle (Te) or even as a coordination conjunction (Cc). The constraint, which assigns the Cs value to the @ana attribute of word form 'че', ignores the morphosyntactic context, but still achieves high enough precision – 99.58 %.

The rules are ordered strictly, since some of them are dependent on the outcome of others. Thus, for example, all rules that apply to elements that usually can be found in the NP internal structure, are preceded by rules that target tokens with possible noun morphosyntactic analysis.

In case there is no intersection between the @aa set of tags and the @svm tag, and the rules do not suggest another solution, the features of the tags from the @aa attribute are compared with the features of the @svm tag. First, the system compares the part-of-speech features of the tags in @aa attribute with the part of speech tag of @svm attribute. If they do not coincide, then the SVMTool decision about the word class of the token is regarded wrong and token ambiguity remains unresolved. If the @svm attribute shares the value of its first position with the first position of one of the alternative tags from the @aa attribute, all other positions are compared in turns. Whenever they do not match, the @aa attribute feature is preferred. In the following

example the verb form can be in different tenses and persons, but always transitive. The suggestion of SVMTool is that the form is intransitive, but the decision that the verb form is in present tense, third person, is correct. In order to repair this problem, the system copies the non-matching information from the @aa attribute to the result in the @ana attribute. In the example, the third position of the tag encodes perfectivity – *p* (perfective) or *i* (imperfective), and the seventh position encodes the tense – *o* (aorist) or *r* (present tense):

```
<ns2:token aa="Vpptf-o2s;Vpptf-o3s;Vpptf-r3s" ana="Vpptf-r3s" svm="Vpitf-r3s">позабавлява</ns2:token>
```

Since the @aa set excludes the perfective verb analysis, the third feature has been rewritten. According to the information encoded in the inflectional dictionary, the value for the present tense for this token (seventh tag position) is possible, so it has not been changed. The percent of ambiguous tokens remaining in the output is about 2.55%. In our future work we will extend the set of rules in order to suggest the most probable solution for these cases.

### 3.3 Lemmatizer

We have implemented a functional lemmatization module, based on the morphological lexicon, mentioned above. The functions are defined via two operations on word forms: *remove* and *concatenate*. The rules have the following form:

```
if tag = Tag then {remove OldEnd; concatenate NewEnd}
```

where Tag is the tag of the word form, *OldEnd* is the string which has to be removed from the end of the word form and *NewEnd* is the string which has to be concatenated to the beginning of the word form in order to produce the lemma. Here is an example of such a rule:

```
if tag = Vpitf-o1s then {remove ox; concatenate a}
```

The application of the rule to the past simple verb form for the verb *чемоx* (remove: *ox*; concatenate: *a*) gives the lemma *чема* (to read). Additionally we encode rules for unknown words like guesser word forms: *#ox* and tag=*Vpitf-o1s*. In these cases the rules are ordered.

In order to facilitate the application of the rules, we attach them to the word forms in the lexicon. In this way, we gain two things: (1) we implement the lemmatization tool as a part of the regular grammar for lexicon look-up, discussed above and (2) the level of ambiguity is less than 2% for the correct tagged word forms. In case of ambiguities we produce all the lemmas. After the morphosyntactic tagging, the rules that correspond to the selected tags, are applied.

### 3.4 Web service architecture and WebLicht compatibility

In the context of eScience researchers want not only to share their resources and technologies, but also to minimize the work needed to reuse them. One of the major current problems is that many of the technologies are incompatible with each other.

Although some have chosen to implement general data-encoding standards like the TEI, many linguistic tools and resources develop their own operational annotation formats. And very few choose to implement common interfaces. These facts impede the interoperability of language technologies. To make sure that the morphological analyzer for Bulgarian can be shared and reused properly, we decided to adopt some of the good ideas of the new-generation Linguistic Resources and Technologies (LRT) project D-SPIN part of the CLARIN Project). Its platform WebLicht is a web-based service environment that allows the users to integrate and use various language resources and tools [4]. The purpose of the platform is to make possible for the scientists to upload their resources and share their tools in one place with common operational and annotation formats, thus improving their collaboration. Although the morphological analyser is not a large scale project and does not intend to produce a web-based environment with quite such capabilities, many of the solutions and ideas that WebLicht provides, are certainly applicable in its context.

The WebLicht platform addresses the two main problems of the research collaboration: different data annotation formats, and technical issues and support of tools/technologies. A common data annotation format allows the resources and tools registered on the platform to be chained together forming flexible linguistic chained processes. Although it is neither required, nor impossible to do otherwise<sup>3</sup>, the use of the Text Corpus Format (TCF), which is a stand-off XML annotation format developed within the D-SPIN project, is recommended. The structure of TCF documents is based on information blocks whose elements are connected by references allowing the annotation of different kinds of information with possible overlapping scope in one file. TCF annotation also makes adding and removing information from the document painless and error-prone, which is a key advantage when putting together custom tools such as the one we described in this paper. The WebLicht architecture requires all the tools to be implemented as RESTful web services that are also recommended to work with the TCF. Web services offer a simple and painless solution to the problem of installing and configuring tools by allowing the authors to host and support them while they are being used in more complex tools. In this way the different steps of a linguistic analysis may be carried out by different tools in different places producing one final result. Adopting the TCF document annotation and implementing all analysis steps as web services not only allows us to share our specific tools through the WebLicht platform, but also to develop, extend and improve our complex tools with less effort.

---

<sup>3</sup> The WebLicht platform allows the registration of web services with any kind of input or output, so in principle other formats could be used. However, tool chains that use such tools should make sure they either include appropriate converters in the chain, or use the same annotation standard in all their tools.

## 4 Error analysis

Here some qualitative error analysis is given for the SVM tagger output as well as the result from the whole morphological pipe.

The errors after the SVM tagger concern either parts of speech, or some of their characteristics. Let us present them.

### Wrong part-of-speech tag

The first case is when the assigned part-of-speech tag is wrong for a *non-ambiguous* word. The following more frequent sub-cases are observed:

- the word is considered a participle instead of a finite verb: ‘*престана да говори*’ (she/he *stopped* speaking);
- the word is considered a noun instead of a participle – ‘*посърнали*’ (haggard-3<sup>rd</sup> person, pl);
- the word is considered a noun instead of an adjective – ‘*сънливи*’ (sleepy-3<sup>rd</sup> person, pl);
- the word is considered an adjective instead of a noun – ‘*капчица любов*’ (‘*droplet* love’, droplet of love). This error is typical for NP phrases of type – NN (or NP NP). This error can be explained by the fact that the position of the first noun in the NP is occupied usually by an adjective.

An interesting case for this automatic tagger is the analysis of the family names. If there is a sequence of a given name and a family name of a person in the sentence, the SVM Tagger would annotate correctly the family name. However, if there is only a family name in the sentence, the SVM Tagger would annotate it like an adjective or a participle (regardless of the capital letter).

The second case is when the assigned part-of-speech tag is wrong for an *ambiguous* word. The following more frequent sub-cases are observed:

- the word is considered an adjective instead of an adverb – ‘*лицето ѝ беше извънредно слабо*’ (her face was *extremely* thin);
- the word is considered an adverb instead of an adjective – ‘*лицето ѝ беше извънредно слабо*’ (her face was *extremely thin*).

It is worth noting that the forms from the above example ‘*извънредно*’ (extreme or extremely) and ‘*слабо*’ (thin or thinly) are POS homonyms in Bulgarian. They can be realized as an adjective or an adverb depending on the context. Other examples for such types of homonyms are: ‘*преди*’ – a preposition (previous to) or an adverb (before); ‘*си*’ – a pronoun (personal or possessive reflexive) or a verb (be-2 person, sg).

### Wrong morphosyntactic features in a tag

The erroneous features often belong to the classes of the verbs and the pronouns. The automatic analysis of the verbs has the following problems to solve:

- whether the verb is transitive or intransitive;
- whether the verb is personal or impersonal;
- whether the aspect of the verb is perfective or imperfective;
- whether the tense is present (1 person, sg) or past simple (2 or 3 person, sg)

These errors can be corrected by the morphological dictionary when the word is not ambiguous. However, the problems remain with the verb aspect.

The analysis of the pronouns is also difficult, because very often one form expresses more than one meaning. For example, the pronoun ‘my’ is homonymous between a personal pronoun (I told *him*) or a possessive pronoun (*his* book).

Another typical error is the wrong gender of a noun. However, this one can be corrected by the information from the morphological dictionary.

Thus, a conclusion can be made that some of the errors of the SVM tagger can be overcome by the application of the dictionary. Other can be addressed by a set of linguistic rules. However, a strategy was needed for a final selection of the correct tag among the competing suggestions (see Section 3.2). Although in many cases the described strategies worked, in some cases they led to a wrong selection.

The accuracy obtained after applying the lexicon look-up and the morphosyntactic context rules is 94.65%. The corrections affect predominantly the morphosyntactically unambiguous tokens – about 3/4 are directly assigned with the only possible POS tag from the lexicon. Worse results have been achieved when the token is ambiguous: 1/3 of the rule-based decisions are correct, 1/5 of the tokens received as a POS tag a subset of the set of possible analyses, or retains their ambiguity (@aa and @ana attributes share their values). As expected, most of the errors stem from the verb forms homonymy.

As the results demonstrate, the combination of a statistical tagger over a rich tagset in combination with linguistic knowledge, encoded within a morphological lexicon and context rules, is as accurate as the machine learning technique over a simplified tagset of [2]. A comparison with [1] is not possible because their tool and the related corpus are not available.

## 5 Conclusion

In this paper we presented a web-enhanced morphological tagger and a lemmatizer for Bulgarian. The work has built on our previous experience with various taggers and tagsets. The processing has been organized in a pipeline, which includes an SVM-based tagger, a look-up of morphological dictionary of Bulgarian and a set of linguistic rules. The decisions behind the pipe have been profiled with respect to

different facts, such as whether the word is morphologically ambiguous, or not; whether the SVM suggested tag is among the tags, encoded in the dictionary. The analyzer is compatible with the Weblicht format, which makes it usable within the CLARIN community, and thus – part of integrated language architectures.

Our next tasks go into two directions: (1) Improvement of the strategy for selecting the correct tags while experimenting with different combinations of processing steps, and (2) Adding a shallow parser on the top of the morphological analyzer and the lemmatizer.

## 6 Acknowledgments

This work has been supported by two European projects: EuroMatrixPlus project (IST-231720) and CLARIN (IST-212230).

## References

- [1] Doychinova, V. and Mihov, S. (2004). High Performance Part-of-Speech Tagging of Bulgarian. In: *Proceedings of Eleventh International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA-2004)*, LNAI #3192, pp. 246-255.
- [2] Georgiev, G., Nakov, P., Osenova, P., Simov, K. (2009). Cross lingual Adaptation as a Baseline: Adapting Maximum Entropy Models to Bulgarian. In: *Proceedings of the Workshop on Adaptation of Language Resources and Technology to New Domains*. In conjunction with RANLP'09, Borovetz, Bulgaria, September 17, 2009. ISBN 978-954-452-009-0. pp 35-38.
- [3] Giménez, J. and Márquez, L. (2004). SVMTool: A general POS tagger generator based on Support Vector Machines. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal.
- [4] Hinrichs, E. W.; Hinrichs, M. & Zastrow, T. (2010) WebLicht: Web-Based LRT Services for German. In *Proceedings of the ACL 2010 System Demonstrations*, pages 25-29, Uppsala, Sweden.
- [5] Joachims, T. Schölkopf, B. (1999). Making Large-Scale SVM Learning Practical. In Burges, C. & Smola, A. (Eds.), *Advances in Kernel Methods – Support Vector Learning*. Cambridge, MA, USA: MIT Press.
- [6] Popov, D., Simov, K., Vidinska, S. (1998). *A Dictionary of Writing, Pronunciation and Punctuation of Bulgarian Language*. (in Bulgarian) Atlantis KL, Sofia, Bulgaria, 927 pp.
- [7] Simov, K. and Osenova, P. (2001). A Hybrid System for MorphoSyntactic Disambiguation in Bulgarian. In: *Proceedings of the RANLP 2001 Conference*, Tzigrav Chark, Bulgaria, 5-7 September 2001. pages 288-290.

- [8] Simov, K., Peev, Z., Kouylekov, M., Simov, A., Dimitrov, M., Kiryakov, A. (2001). CLaRK – an XML-based System for Corpora Development. In: Proc. of the Corpus Linguistics 2001 Conference. Lancaster, UK.
- [9] Simov, K., Osenova, P., Kolkovska, S., Balabanova, E., Doikoff, D. (2004). A Language Resources Infrastructure for Bulgarian. In: *Proceedings of LREC 2004*, Lisbon, Portugal. 2004. pages 1685-1688.
- [10] Vapnik, V. N. (1999). *The nature of statistical learning theory* (2<sup>nd</sup> ed.). New York: Springer.

# Automatic Extraction of Multiword Expressions Using Linguistic Constraints for Slovak LVCSR

Ján Staš<sup>1</sup>, Daniel Hládek<sup>1</sup>, Marián Trnka<sup>2</sup> and Jozef Juhár<sup>1</sup>

<sup>1</sup> Dept. of Electronics and Multimedia Communications, FEI,  
Technical University of Košice, Park Komenského 13, 042 00 Košice, Slovakia  
{jan.stas, daniel.hladek, jozef.juhar}@tuke.sk

<sup>2</sup> Department of Speech Analysis and Synthesis, Institute of Informatics,  
Slovak Academy of Sciences, Dúbravská cesta 9, 845 07 Bratislava, Slovakia  
trnka@savba.sk

**Abstract.** The most common mistakes of the automatic speech recognition systems arise at the beginning of the speech or after long pause and are caused by misrecognition of short monosyllabic words. To avoid this problem it is suitable to model these events using multiword expressions. This article aims to describe a process of automatic extraction of the most frequent multiword expressions from the text corpora that are suitable for inclusion into the model of the Slovak language in order to reduce mistakes in recognition of the short words. The automatic extraction of the multiword expressions uses standard statistical measures that are based on the co-occurrence of these word pairs in the text corpora limited by the linguistic constraints. Experimental results show moderate improvement in the recognition accuracy of the short monosyllabic words in Slovak LVCSR task.

## 1 Introduction

Multiword expression (MWE) can be described as a lexical item that can be decomposed into multiple simplex word and display *lexical, syntactic, semantic, pragmatic* or *statistical* meaning or idiosyncrasy [8]. It means that the MWEs can be characterized as a sequence of at least two or more words that co-occur together frequently in the given language. They are usually associated with a fixed set of situations or particular context and the syntax of the MWE is neither derived directly from that of its components nor explicitly or implicitly derivable from its parts [4].

MWEs have an important role in the field of the computational linguistics (CL) and natural language processing (NLP) and they can be formed as *phrasal verbs, light verbs, idioms, nouns* or *nominal compounds, institutionalized phrases* and some other *collocations* that co-occur much more often than would be expected by chance. MWEs are often used in *multi- or cross-lingual information retrieval* systems for word-alignment tasks, in *information extraction, querying* or *question answering* or in *machine translation* [7].

Nowadays, several research works have been focused on the possibility of use MWEs in the language modeling to improve the accuracy of an automatic speech recognition (ASR) system. As it was published in [5] on including MWEs in language model of the Czech language, misrecognition of the short words can be effectively avoided and overall precision of the ASR system can be enhanced.

In order to be able to model any language using MWEs, it is necessary to extract relevant multiwords from the text corpora used in the process of training language models.

The selection of the MWE in that manner depends on the type of the text data and the field of application. The most common methods of MWE extraction from the text corpus can be divided into following groups [1]: *a. statistical methods*, which use association measures to rank MWE candidates; *b. linguistic methods*, based on morpho-syntactic patterns; *c. hybrid methods*, making use both previous statistical measures and linguistic filters and *d. word alignment* models. For the automatic extraction of MWEs in the Slovak language for the purpose of their utilization in the *large vocabulary continuous speech recognition* (LVCSR) system, the proposed approach is focused on the hybrid method where statistical measures are limited by the linguistic constraints.

This article is organized as follows. In Section 2, a short overview about the problem of language modeling, some common mistakes in speech recognition in Slovak language and their solution are described. The text corpora used for automatic extraction of relevant MWEs is described in Section 3. Several standard measures for extraction of the relevant MWEs which were applied in our task and some linguistic constraints that were designed are presented in Section 4. In Section 5 the LVCSR setup for performing experiments is briefly described. In the following sections the experimental results about proposed methodology of MWE extraction and their effect on accuracy of speech recognition are summarized. Section 7 concludes this paper.

## 2 Motivation

As it was mentioned in the Introduction of this article, the most common mistakes in the speech recognition occur at the beginning of the speech or after long pause in recognizing of the short monosyllabic words that consist of at most than three or four characters. These words are often added to the following or preceding word, recognized as a noise or ignored [5], what consequently leads to the decreasing of the accuracy of the LVCSR system. It has been showed that MWEs in the form of connection of short (monosyllabic) word with long (di-, tri- or polysyllable) word, which is usually more recognizable, can help increasing the recognition accuracy of the given short word. Moreover, using MWEs increases the order of  $n$ -gram language model and decreases the number of pronunciation variants that depend on the context of the given word because in an inflective language some of the words are pronounced differently in different context [5].

The extraction of MWEs in the Slovak language is performed by following selection criteria: *a.* MWE consists of at least one monosyllable word, that consists of more than three characters; *b.* both words forming the MWE and MWE itself must occur frequently in the language; *c.* final selection is conditioned by additional linguistic constraints.

## 3 Text corpora

In process of creation an effective model for any inflective language, it is important to collect a large amount of text data that enter to the process of training language model (LM). These text corpora were created using an automatic system designed in our laboratory that retrieves the text data from various Internet pages and electronic sources that are written in the Slovak language [3]. The text data were normalized by additional modifications as word tokenization, sentences segmentation, abbreviations expanding, numerals transcription, etc. System also includes constraints such as filtration of grammatically

incorrect words, duplicity verification of text documents and others [9]. Processed text corpora were divided into four domain-specific corpora. In the Table 3 it can be seen the statistics of the number of words and sentences for particular text corpus.

text corpus	# words (W)	# sentences (S)	AVG (W/S)
WEB CORPUS	948 958 508	54 765 873	17.328
BROADCAST NEWS	590 274 484	33 804 173	17.462
LEGAL CORPUS	258 131 635	9 135 908	28.255
OTHER TEXTS	57 169 452	3 899 675	14.660
HELD-OUT DATA	55 163 941	1 782 333	30.950
<b>TOTAL</b>	<b>1 909 698 020</b>	<b>103 387 962</b>	<b>18.471</b>

**Table 1.** Statistics on text corpora

A vocabulary have been generated from each corpus, merged, passed through the spellcheck lexicon and then was manually checked and corrected.

We did not use the text corpora from the Slovak National Corpus database by reason that the text data contained in this database are not normalized and the text style do not cover the area of speech recognition in realized experiments.

type of MWE	WEB	B. NEWS	LEGAL	OTHER	AVG
two long words	65.226 %	65.490 %	64.607 %	68.038 %	65.840 %
one short word at least	24.198 %	24.407 %	22.857 %	22.745 %	23.552 %
numerals or punctuation	10.577 %	10.103 %	12.536 %	9.217 %	10.608 %

**Table 2.** Representation of MWEs in bigram statistics for particular text corpus

From each corpus, bigram counts were generated, representing the input of the system for automatic extraction of multiwords, described in Section 4. Note that among selected MWEs are not those that contain at least one ordinal or cardinal number or punctuation (see Table 3).

## 4 Automatic extraction of multiword expressions

For the extraction of the multiword expressions from bigram counts obtained from the text corpora, three common measures have been used. The measures are based on the standard statistical functions of the *absolute and relative frequency* in the corpus and *pointwise mutual information* that are described below. Obtained MWE were ranked according to their occurrence in corpus and then additionally delimited by the selected linguistic constraints.

### 4.1 Statistical measures for extracting relevant MWEs

- **Absolute frequency** of the MWE ( $f_A$ ) expresses the total number of occurrences of the MWE in the whole corpus.

- **Relative frequency** of the MWE in the context ( $f_R$ ) indicates the percentage of all occurrences of the word  $y$  in the corpus is in the context of the word  $x$ .

$$f_R(x, y) = \frac{f_A(x, y)}{f_A(x)} \times 100\%. \quad (1)$$

- **Pointwise mutual information** (PMI) is a measure of how much the actual probability of a particular co-occurrence of events  $p(x, y)$  differs from what we would expect it to be on the basis of the probabilities of the individual events and the assumption of independence  $p(x)p(y)$ .

$$PMI(x, y) = \log_2 \frac{p(x, y)}{p(x)p(y)}. \quad (2)$$

Selection these measures was intentional. In general, absolute frequency expresses the most frequent events in given language. Using relative frequency in the context of the first word, we can effectively extract MWEs such part-of-speech in the Slovak language as prepositions, conjunctions or pronouns that usually occur in the first place of given MWE. Finally, PMI reflects such multiwords as collocations that do not occur in language frequently but usually have certain meaning.

## 4.2 Linguistic constraints

Linguistic constraints come from the observations of the behaviour of a LVCSR system in the process of testing on the different types of test data. It have been discovered that our LVCSR system is often wrong in following cases. In the first case, there was an *assimilation of voicing* (voice or voiceless) on a word boundaries. In the second common case, the system was wrong if a first word in MWE ended with *same letter* as the second word begins (see Table 4.2). Finding out these simple rules have enabled a possibility to significantly reduce the amount of extracted relevant MWEs from text corpora using statistical measures.

absolute frequency $f_A$		relative frequency $f_R$		pointwise mutual inf. $PMI$	
assimilation of voicing	similarity in letters	assimilation of voicing	similarity in letters	assimilation of voicing	similarity in letters
v právnej	mu ukladá	v súčasnosti	po odtrhnutí	žuť žuvačku	húf fanúšičiek
keď sa	len na	k revidovaniu	do ohnivej	očistných kúr	skrutkového oja
z toho	od dlžníka	s difrakciou	bez zapnutých	rys ostrovid	popadalo ono
v paragrafe	do omeškania	k sebazáchove	súd dedukoval	hus zagágala	poškriabal lak
ak by	som mal	pod tatrami	ju uviazať	krčných žíl	potopil loď

**Table 3.** Examples of extracted the most frequented MWEs in Slovak language

For each statistical measure, first 10 000 word-pairs from each text corpus have been selected. After the selection, in order to avoid the out-of-domain effect, resulting MWEs for each individual measure were represented by a weighted (average) sum of MWEs from each corpus. After merging and manually correction of the individual lists with MWEs, the dictionary with unique 3 000 multiwords have been obtained. Selected MWEs came mostly from the statistics of the absolute and relative occurrences, because the most common

mistakes in recognition are more probable in these cases than in cases selected by the PMI measure. Finally, dictionary entering into the recognition process by LVCSR system have been extended by multiwords with manually checked phonetic transcription.

It should be noted that the lemmatisation plays an important role in automatic extraction of MWEs in any Slavic language. By reason of absence of any tool for lemmatisation the Slovak language, we did not use lemmatisation and relied only on proposed methodology of extraction MWEs from our text corpora.

## 5 LVCSR setup

The text corpora have been processed to include multiwords into the process of the training LMs. Then, *the SRI Language Modeling Toolkit* [11] has been used for building LMs from particular text corpora. Trigram LMs were created with vocabulary size of 351 473 unique words (includes 3 000 MWEs) and smoothed by using *the modified Kneser-Ney algorithm*. Obviously, the unigram counts of words forming given MWE have been increased by the bigram count of MWE in given text corpora by reason of alignment the statistics of these words in training process. Then particular trigram LMs were adapted to the domain of judicature and combined by *linear interpolation*, where interpolation weights were set using EM algorithm computed on the held-out data set [10]. The same training process was performed for LM that does not include MWEs.

The triphone context-dependent acoustic model based on the *hidden Markov models* (HMM) has been used, where each state have been modeled by 32 Gaussian mixtures. The model has been generated from feature vectors that contains 39 *mel-frequency cepstral* (MFC) *coefficients*. It has been trained using about 250 hours of gender balanced read speech recordings of judicial proceedings, recorded from 250 speakers and 100 hours of spontaneous speech, recorded from 120 speakers at council hall. For acoustic modeling rare triphones *the effective rule-based triphone mapping algorithm* was used [2]. The training set also involves model of a silence, short pause and additional noise events.

For speech the recognition, *the LVCSR engine* [6] with recognition algorithm based on the two-pass strategy has been used. The input data are processed in the first pass with bigram LM, and the final search for trigram model is performed again using the result of the first pass to narrow the search space.

The test data were represented by 20 hours of recordings obtained by randomly selected segments from each speaker contained in first read speech database that were not used in training and contain 41 878 words in 3 426 sentences and phrases. We have also used phrases in our test set, because in real conditions, people do not only break on sentence boundaries but also on phrase boundaries.

Two standard measures for evaluation of the LM based on *word error rate* (WER) and *perplexity* (PPL) at the test set have been used. WER is a standard measure of the performance of the ASR system, computed by comparing reference text read by a speaker against the recognized result and takes into account insertion, deletion and substitution errors. If the ASR system is not available, the perplexity is often used for evaluation. It is defined as the reciprocal of the (geometric) average probability assigned by the LM to each word in the test set.

## 6 Experimental results

The experiments with MWEs were oriented on the evaluation WER and PPL on the test set to discover their effect on the overall recognition accuracy of the LVCSR system as well as the impact of this concept on the particular hypotheses after speech recognition. The LM with multiwords were compared with the standard word-based LM, where both have been trained in same manner.

language model	# bigrams	# trigrams	xRT*	PPL	WER [%]
word-based LM	81 974 817	109 574 999	3.64	57.2970	6.26
LM with MWEs	85 245 932	122 435 417	3.33	62.7111	6.22

\*real-time factor represents the ratio between computational time of the process of speech recognition and real duration of speech recordings.

**Table 4.** The comparison of LMs in WER and PPL on the test set.

As we can see in the Table 6, only small number of MWEs (3 000 MWEs) caused a moderate increase in number of bigrams and trigrams in given LM. This knowledge has also connection with increase in perplexity of the LM, approximately 9.45% relative. The increase in PPL is caused by the extension of vocabulary with MWEs and undesired shortening of history of  $n$ -grams contained in MWEs that results in decreasing of predictive ability of that LM. On the other hand, we have observed a slight decrease in WER, approximately 0.64% relative. There is no significant result, but implies the usability of this concept in speech recognition. Therefore, MWEs should be selected considering the most frequented errors after speech recognition and used in the next training of the LM. Besides, multiwords make the process of recognition faster (see xRT in Table 6). It is because sentences that contain multiword expressions are composed from a smaller number of words than sentences without multiwords.

language model	in all MWEs	in error cases
word-based LM	11.49%	59.18%
LM with MWEs	7.93%	40.82%
joint errors	5.41%	27.82%

**Table 5.** Representation of errors in hypotheses after recognition only for MWEs.

As it has been stated above, second experiment has been oriented on discovering influence of MWEs on individual hypotheses after recognition. In Table 6 we can see that LM with MWEs has outperformed the standard word-based LM too in number of errors in recognized hypotheses (approximately 20% less), especially at the beginning of sentences or after a long pause. First column in the table (see “in all MWEs”) denoted the representation of errors among all 915 MWEs that occur in the test set. Also, second column (see “in error cases”) represents error rates among all wrong recognised MWEs.

Joint errors represent mismatch in hypotheses of multiwords of both LMs after recognition against reference MWEs. It should be noted that a number of errors in recognition could be caused by still insufficient phonetic restrictions and transcription pronunciation between words forming MWE in dictionary. This information will be a subject of further research in modeling Slovak language for the LVCSR system.

## 7 Conclusion

In this article, the process of extraction and modeling Slovak language using multiword expressions has been presented. Selection of relevant MWEs was based on standard statistical measures for extraction multiwords in systems for natural language processing and was delimited by several linguistic constraints. Modeling Slovak language using MWEs brought a slight increase of the recognition accuracy and decreasing errors in recognition of the short words at the beginning of sentences or after long pause. The improvement in recognition accuracy using LM with MWEs could be achieved by precise phonetic restrictions and transcription pronunciation between words forming MWEs in dictionary and selection these units from specific area of speech recognition. Further research should be also focused on the elimination of these errors to achieve a better performance of the Slovak LVCSR system.

## Acknowledgment

The research presented in this paper was supported by the Slovak Research and Development Agency and Ministry of Education under research project VMSP-P-0004-09 and VEGA-1/0065/10.

## References

- [1] Attia, M., Toral, A., Tounsi, L., Pecina, P., and van Genabith, J. (2010). Automatic Extraction of Arabic Multiword Expressions. In *Proc. of the COLING Workshop of Multiword Expressions: from Theory to Applications (MWE 2010)*, pages 18–26, Beijing, China.
- [2] Darjaa, S., Cerňak, M., Š. Beňuš, Rusko, M., Sabo, R., and Trnka, M. (2011). Rule-Based Triphone Mapping for Acoustic Modeling in Automatic Speech Recognition. *I. Habernal and V. Matoušek (Eds.): TSD 2011, Springer-Verlag, LNAI 6836*, pages 268–275.
- [3] Hládek, D. and Staš, J. (2010). Text mining and processing for corpora creation in Slovak language. *Journal of Computer Science and Control Systems*, 3(1):65–68.
- [4] Kim, S. N. (2008). *Statistical Modeling of Multiword Expressions*. PhD thesis, Dept. of Computer Science and Software Engineering, University of Melbourne, Australia.
- [5] Kolorenč, J., Nouza, J., and Červa, P. (2006). Multi-words in the Czech TV/Radio News Transcription System. In *Proc. of the 11th Intl. Conf. Speech and Computer (SPECOM 2006)*, pages 70–74, St. Petersburg, Russia.

- [6] Lee, T., Kawahara, T., and Shikano, K. (2001). Julius – An Open Source Real-time Large Vocabulary Recognition Engine. In *Proc. of EUROSPEECH 2001*, pages 1691–1694.
- [7] Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.
- [8] Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In *Proc. of the 3rd Intl. Conf. on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15, Mexico City, Mexico.
- [9] Staš, J., Hládek, D., Pleva, M., and Juhár, J. (2010a). Slovak Language Model from Internet Text Data. *A. Esposito et al. (Eds.): Towards Autonomous, Adaptive, and Context-Aware Multimodal Interfaces: Theoretical and Practical Issues, Springer-Verlag, LNCS 6456*, pages 352–358.
- [10] Staš, J., Trnka, M., Hládek, D., and Juhár, J. (2010b). Text Preprocessing and Language Modeling for Domain-Specific Task of Slovak LVCSR. In *Proc. of the 7th Intl. Conf. on Digital Technologies, Circuits, Systems and Signal Processing (DT 2010)*, pages 1–5, Žilina, Slovak Republic.
- [11] Stolcke, A. (2002). SRILM – An Extensible Language Modeling Toolkit. In *Proc. of the 7th Intl. Conf. on Spoken Language Processing (ICSLP 2002)*, pages 901–904, Denver, Colorado.

# Common Formal Framework for Multilingual Representation of Inflectional Morphology for Two Related Slavonic Languages

Velislava Stoykova

Institute for Bulgarian Language, Bulgarian Academy of Sciences,  
52, Shipchensky proh. str., bl. 17, 1113 Sofia, Bulgaria  
vili1@bas.bg

**Abstract.** The paper analyzes the basic assumptions for presenting nominal inflectional morphology using DATR language for lexical knowledge presentation for Russian and Bulgarian language. It compares the principles and motivation of the proposed encodings, which use non-monotonic orthogonal semantic networks. Finally, a more general principles for presenting inflectional morphology are offered by introducing a semantic hierarchy using traditional grammar features.

**Keywords:** Natural Language Processing, knowledge representation, DATR language for lexical knowledge presentation, computational morphology, semantic networks.

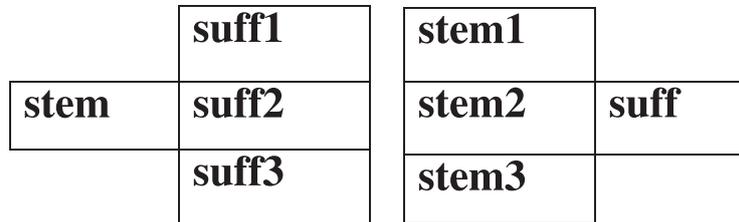
## 1 Introduction

Slavonic languages have had a long parallel historical development and as a result they share similar grammar features at the level of phonetics, morphology, and syntax. They are widely known as languages which use cases to represent syntactic structures. At the same time, a great deal of formal syntactic theories have been developed, predominantly for English, which do not offer a sufficiently good morphological account. The problem is that before starting the syntactic analysis, a morphological interpretation have to be developed. It is more complicated when a morpho-syntactic phenomena have to be interpreted.

Another open question is whether the principles of formal interpretation of one Slavonic language are valid for the others. Further, we are going to analyse the formal interpretation of nominal inflectional morphology for Russian and Bulgarian with respect to multilingual applications.

## 2 The traditional academic representation and computational morphology formal models of inflectional morphology

The traditional interpretation of inflectional morphology given at the academic descriptive grammar works [8] is a presentation of tables. The tables consist of all possible inflected forms of a related word with respect to its subsequent grammar features. The artificial intelligence (AI) techniques offer a computationally tractable encoding preceded by a related semantic analysis, which suggest a subsequent architecture. Representing inflectional morphology in AI frameworks is, in fact, to represent a specific type of grammar knowledge.



**Fig. 1.** Conjugational solution and variant stem solution.

The standard computational approach to both derivational and inflectional morphology is to represent words as a rule-based concatenation of morphemes, and the main task is to construct relevant rules for their combinations. With respect to the number and types of morphemes, the different theories offer different approaches depending on variations of either stems or suffixes as follows:

- (i) Conjugational solution offers invariant stem and variant suffixes, and
- (ii) Variant stem solution offers variant stems and invariant suffix.

Both these approaches (Fig.1) are suitable for languages, which use inflection rarely to express syntactic structures, whereas for those using rich inflection some cases where phonological alternations appear both in stem and in concatenating morpheme a “mixed” approach is used to account for the complexity. Also, some complicated cases where both prefixes and suffixes have to be processed require such approach.

We evaluate the “mixed” approach as a most appropriate for the task because it considers both stems and suffixes as variables and, also, can account for the specific phonetic alternations. The additional requirement is that during the process of inflection all generated inflected rules (both using prefixes and suffixes) have to produce more than one type of inflected forms.

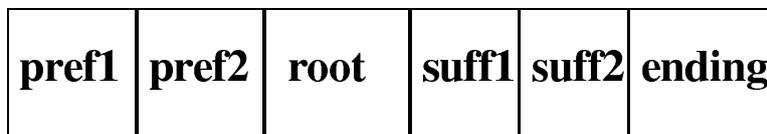
We evaluate the DATR language for lexical knowledge presentation as a suitable formal framework for analyzing and presenting Slavonic nominal inflectional morphology.

## 2.1 Interpreting sound alternations

Natural language processing applications use different techniques to represent and differentiate between phonological, morphological, and syntactic knowledge even all these types represent the inherent language features.

The problem of interpreting sound alternations of inflectional morphology is a key problem of any formal presentation of Slavonic languages. The sound alternations influence the inflectional morphology of almost all part-of-speech and as a result they form irregular word forms. In fact, we have a rather unsystematically formed variety of regular and irregular sound alternations which is very difficult to be interpreted formally.

The phonetic alternations are of various types and influence both derivational and inflectional morphology. The classification accepted in the academic descriptive grammar works does not concern the formal way the sound alternations could be interpreted.



**Fig. 2.** The word structure according to the general linguistic morphological theory.

The general morphological theory offers a segmentation of words (Fig. 2) which consists of root to which prefixes, suffixes or endings are attached. Normally, all three types of morphemes are used and additional difficulties come from the fact that sound alternations can be occurred both in stems, prefixes, suffixes, and also on their boundaries which suggest extremely complicated solutions.

The formal account of sound alternations have to be offered only for those sound alternations which are significant for the inflectional morphology. It is possible to be done by using both the architecture of the application (the definition of inflectional rules) and the lexical information presentation scheme. Also, it requires the interpretation of both phonetic and morphological features in one consistent formal representation.

We evaluate the DATR language for lexical knowledge presentation as a suitable formal framework for analyzing and presenting Slavonic nominal inflectional morphology.

### 3 The DATR language

The DATR language for lexical knowledge presentation is a non-monotonic language for defining the inheritance networks through path/value equations [7]. It has both an explicit declarative semantics and an explicit theory of inference allowing efficient implementation, and at the same time, it has the necessary expressive power to encode the lexical entries presupposed by the work in the unification grammar tradition [4, 5, 6].

In DATR information is organized as a network of nodes, where a node is a collection of related information. Each node has associated with it a set of equations that define partial functions from paths to values where paths and values are both sequences of atoms. Atoms in paths are sometimes referred to as attributes.

DATR is functional, it defines a mapping which assigns unique values to node attribute-path pair, and the recovery of these values is deterministic. With respect to its universality, DATR's formal properties and techniques underlay both the rule-based inference and non-monotonic inference by default, and allow to account for language phenomena such as regularity, irregularity, and subregularity by using deterministic parsing.

The semantics of DATR uses non-monotonic inference and default inheritance, and allows the generalization-capturing representation of the inflectional morphology. DATR has the expressive power which is capable to encode and process both syntactic and morphological rules and it allows representation of grammar knowledge by using the semantic networks.

The DATR language for lexical knowledge presentation offers syntagmatic operators (which can be used to define the concatenation procedures) and paradigmatic operators (which can be used to define the specific structure of the inflecting rules, and for further development of the part-of-speech interpretations).

The DATR language has been used for developing inflectional morphology for a lot of languages including Slavonic languages. It has a lot of implementations, however, the encodings of the analyzed applications are presented by various files available at <http://www.cogs.susx.ac.uk/lab/nlp/datr/datrnode49.html> for related languages. This PROLOG encoding uses Sussex DATR notation [12].

DATR allows construction of various types of language models (language theories), however, the analyzed interpretations underlay the architecture of a rule-based formal grammar and a lexical database. The particular queries to be evaluated are all related inflected word forms, and the implementation allows to process words in Cyrillic alphabet.

#### 4 Russian nominal inflectional morphology in DATR

The DATR language has been applied for developing inflectional morphology for a lot of languages including Russian [3]. The ideas used for Russian nominal inflection interpretation offered by Corbett and Fraser underlay that of a paradigm and the encoding presents a resolving of a tabular conceptualization encoding task.

In fact, the Network Morphology is a framework for describing inflection which offers a formally explicit account of lexical entries, declensional classes, word classes, and the relationships between them by giving a set of universal constraining principles of morphology.

The analyzed application of inflectional morphology of Russian nominal inflection is linguistically motivated. In particular, the underlying basic idea of the analysis is to reconsider the Russian declensional classes described in Zaliznjak's dictionary, however, the approach adopted has implications well beyond the Russian. The interpretation uses declensional classes, i.e. the Word and the Paradigm framework and the features of case, number, and animacy as a starting point of the formal analysis, which is of theoretical value since it presents four declensional classes instead of three, presented traditionally. It consists of a formal grammar (inflectional rules) and a lexical database (nouns of all declensional classes) and the queries to be evaluated are all inflected word forms.

Further, we are going to analyse the fragment of encoding presenting the Russian nouns inflection for the features of case and number [3].

It starts with the node NOMINAL which interprets the inflectional rules which are defined with respect to the phonetical sound alternations by using the inflectional phonetical and morphological stems depending on their hardness for both nouns and adjectives.

The node NOMINAL is as follows <sup>1</sup>:

```
NOMINAL:
<stem> == "<infl_root>"
<phon stem hardness> == hard
<mor stem hardness> == "<phon stem hardness>"
<acc> == "<mor nom>"
<acc pl animate> == "<mor gen pl>"
<acc sg animate masc> == "<mor gen sg>"
```

<sup>1</sup> Here and elsewhere in the description we use Latin alphabet to present morphemes instead of Cyrillic used normally. Because of the mismatching between both some of the typically Bulgarian phonological alternations are assigned by two letters, whereas in Cyrillic alphabet they are marked by one.

```

<mor acc $number> == <acc $number "<syn animacy>" "<syn gender>">
<mor dat pl> == "<stem pl>" "<mor theme_vowel>" _m
<mor inst pl> == "<stem pl>" "<mor theme_vowel>" _m'i
<mor loc pl> == "<stem pl>" "<mor theme_vowel>" _x.

```

The node GENDER is introduced to differentiate between different types of gender assignment (including the semantic gender defined as 'formal') and is as follows:

```

GENDER:
<male> == masc
<female> == fem
<undifferentiated> == "<formal gender>".

```

The basic node which defines the general rules of nouns inflection is the node NOUN. It inherits the grammar rules of node NOMINAL but also defines new inflectional rules.

```

NOUN:
<> == NOMINAL
<mor loc sg> == "<stem sg>" _e
<mor nom pl> == "<stem pl>" _i
<mor gen pl> == "<<mor stem hardness>" mor gen pl>"
<soft mor gen pl> == "<stem pl>" _ej
<mor theme_vowel> == _a
<syn cat> == n
<syn animacy> == "<sem animacy>"
<syn gender> == GENDER: <<sem sex>">
<sem sex> == undifferentiated.

```

The node N\_0 defines nouns which are assigned to declensional types I and IV and it inherits all grammar rules from node NOMINAL but introduces new inflectional rules.

```

N_0:
<> == NOUN
<mor gen sg> == "<stem sg>" _a
<mor dat sg> == "<stem sg>" _u
<mor inst sg> == "<stem sg>" _om.

```

Node N\_I defines nouns which belong to I declension. It inherits all inflectional rules from node N\_0 and introduces new inflectional rules.

```

N_I:
<> == N_0
<formal gender> == masc
<mor nom sg> == "<stem sg>"
<hard mor gen pl> == "<stem pl>" _ov.

```

The example Russian word for law 'zakon' which uses the inflectional rules of node N\_I is defined as a separate node through the <infl\_root> and <sem animacy>. The generated case inflected word forms are given at the Appendix.

```
Zakon:
<> == N_I
<infl_root> == zakon
<sem animacy> == inanimate.
```

The entire application of nominal inflectional morphology uses new insights into specific areas of Russian inflectional morphology like paradigm, gender assignment, case, number, and animacy. It presents the declensional classes as nodes of inheritance hierarchy and uses default inheritance hierarchy to model word structure by using a great deal of information sharing. It represents the inflectional morphology as a network of hierarchies and differentiates between the lexemic hierarchy and the inflectional hierarchy by using semantic principles.

## 5 Bulgarian nominal inflectional morphology in DATR

The standard Bulgarian language does not use cases for syntactic representation but it has very rich inflectional system [8]. Also, it uses prepositions and a base word form instead of case declensions. It is considered as a language which use relatively free word order, so the subject can take every syntactic position in the sentence (including the last one). Another important grammar feature of Bulgarian is the feature of definite article which is an ending morpheme [8]. The fact gives a priority to morphological interpretations of definiteness in spite of syntactic since at the level of syntax, the definite article shows the subject (when it is not a proper name).

### 5.1 The semantics of definiteness and its formal morphological marker

The definiteness in Bulgarian may express various types of semantic relationships like a case (to show subject), part-of-whole, deixis etc. The definite article can assign an individual or quantity definiteness, and it has a generic use as well.

The syntactic function of definiteness in Bulgarian is expressed by a formal morphological marker which is an ending morpheme [8]. It is different for genders, however, for the masculine gender two types of definite morphemes exist to determine a defined in a different way entities, which have two phonetic alternations, respectively.

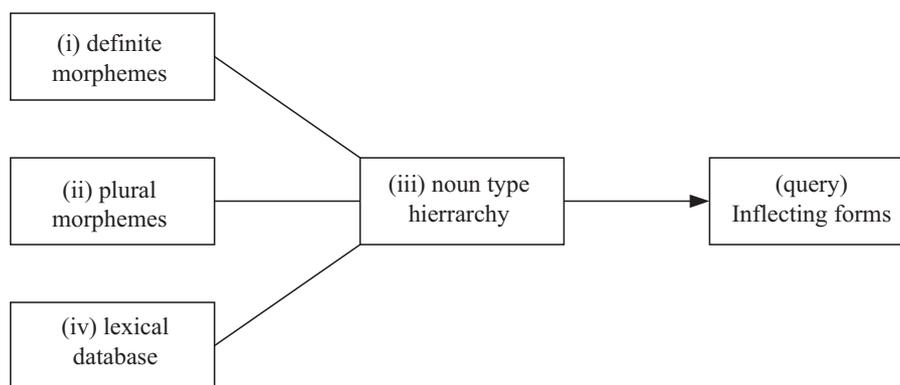
For the feminine and for the neuter gender only one definite morpheme exists, respectively. For the plural, two definite morphemes are used depending on the ending vocal of the main plural form. The following part-of-speech in Bulgarian take the definite article: nouns, adjectives, numerals (both cardinals and ordinals), possessive pronouns (the full forms), and reflexive-possessive pronoun (its full form). The definite morphemes are the same for all part-of-speech, however, in further description we are going to analyze only some general types of rules used for the interpretation of nominal inflectional morphology of definiteness in Bulgarian given in Stoykova [9], [10], and [11].

### 5.2 The architecture of the application

The analyzed application of nominal inflectional morphology of Bulgarian is linguistically motivated. In particular, the underlying basic idea is that of a paradigm since morphemes

are defined to be of semantic value and are considered as a realization of a specific morphosyntactic phenomenon. The words are encoded by introducing different roots to account for the related phonetic alternations, which are defined to be of semantic value as well. The approach is related and indebted to that of Cahill and Gazdar [1, 2] used to account for German nominal inflectional morphology.

The architecture represents an inheritance network consisting of various nodes which allows to account for all related inflected word forms within the framework of one grammar theory. Thus, the general architecture of the application is as follows (Fig. 3):



**Fig. 3.** The general architecture of the model.

(i) All definite inflecting morphemes for all forms of definite article attached to node DET and defined by paths <masC>, <masC\_1>, <femN>, <neut>, and <plur>.

(ii) 12 inflecting morphemes for generating plural forms defined at node Suff.

(iii) The inflectional rules defined as concatenations of morphemes for generation of all possible inflected word forms attached to their related inflectional types nodes.

(iv) The words are given as lexical database attached to their inflectional type nodes. They are defined as lexical entries through paths <root> and <root plur>, so to account for the different phonological alternations.

The DATR logical representation framework uses rule-based reasoning with non-monotonic inference and default inheritance to represent the inflectional rules in semantic network. It suggests the structure of semantic network that can employ the generalization capturing rules in which the grammar knowledge is encoded by the attachment of inflectional rules to the related nodes.

In principle, DATR permits multiple default inheritance and prioritized inheritance enforced by orthogonal representation, and suggest the lexicon being structured mostly by inheritance. This technique allows to account for the grammar irregularities and to use the compilation rules which can generate all possible inflected forms within one application. The application uses a hierarchical structure of the lexical representation in which the feature of gender is a trigger to change the values of the inflected forms.

During the process of inflection, also, various phonetic alternations are taking place. The phonetic alternations at the morpheme boundary are interpreted either by defining

new grammar rules or new nodes, and the phonetic alternations inside morphemes are interpreted by introducing different roots. It is possible, also, to use the technique of finite state transducers [10]. The analysed application interprets, also, more complicated cases of inflection, where both prefixes and suffixes can be processed by defining new nodes of the network.

The encoding is already published [9] and the complete noun inflectional type hierarchy as well. Further, we are going to analyse the fragment of noun inflectional encoding and particularly the core node, which defines the general inflectional rules of the noun inflectional type hierarchy.

### 5.3 The inflectional rules

The DATR analysis of nouns [9, 11] starts with node DET which defines all inflecting morphemes for the definite article and is as follows:

```
DET:
  <sing undef>      ==
  <sing def_2 masc> == _ja
  <sing def_2 masc_1> == _a
  <sing def_1 masc> == _jat
  <sing def_1 masc_1> == _ut
  <sing def_1 femn> == _ta
  <sing def_1 neut> == _to
  <plur undef>      ==
  <plur def_1>     == _te.
```

The node Suff defines 12 inflecting morphemes for generating the plural inflected forms.

```
Suff:
  <suff_11> == _i
  <suff_111> == _ovci
  <suff_12> == _e
  <suff_121> == _ove
  <suff_122> == _eve
  <suff_123> == _ovce
  <suff_21> == _a
  <suff_22> == _ja
  <suff_211> == _ishta
  <suff_212> == _ta
  <suff_213> == _ena
  <suff_214> == _esa.
```

The basic node of noun inflectional types hierarchy is the node Noun and it defines the general inflectional rules for compilation of all possible inflected forms.

```
Noun:
  <suff> == suff_11
  <gender> == masc_1
  <> == <stem> DET: <Idem "<gender>">
  <stem sing> == "<root sing>"
  <stem plur> == "<root plur>"Suff:<"<suff>">.
```

The others word inflectional types are defined either by changing the values of <gender> (the definite morphemes) and <su<sub>f</sub>f> (the plural morphemes) or by introducing new inflectional rules. The inflectional word forms of the example Bulgarian word for law 'zakon' which use the inflectional rules of node Noun are given at the Appendix.

## 6 Conclusions

The analyzed applications of Slavonic nominal inflection offer new insights into specific well-established problem areas of inflectional morphology. They use the traditional grammar features of declension, gender, number, and animacy to encode the features of case and definiteness.

The applications introduce the inheritance hierarchies for concise encoding and represent the declensional classes as nodes. The architecture of the interpretations differentiate between inflectional classes (types) hierarchy, lexemic hierarchy, and semantic hierarchy. It underlay the idea that related languages because of the fact that they share similar grammar features can be formally presented by using similar ideas and techniques for the encoding.

## References

- [1] Cahill L. and Gazdar G. 1999a. German noun inflection. *Journal of Linguistics* 35.1, 1-42.
- [2] Cahill L. and Gazdar G. 1999b. The Polylex architecture: multilingual lexicons for related languages. *Traitement automatique des langues*, 40.
- [3] Corbett, G. and Fraser, N. 1993. Network Morphology: a DATR account of Russian nominal inflection. *Journal of Linguistics*, 29, 113–142.
- [4] Evans R. and Gazdar G. 1989a. Inference in DATR. *Fourth Conference of the European Chapter of the Association for Computational Linguistics*, 66–71.
- [5] Evans R. and Gazdar G. 1989b. The semantics of DATR. In Anthony G. Cohn, ed. *Proceedings of the Seventh Conference of the Society for the Study of Artificial Intelligence and Simulation of Behaviour*, London: Pitman/Morgan Kaufmann, 79–87.
- [6] Evans R. and Gazdar G. 1990. The DATR papers. *CSRP 139, Research Report, vol. 1*, University of Sussex, Brighton.
- [7] Evans R. and Gazdar G. 1996. DATR: A language for lexical knowledge representation. *Computational Linguistics*, 22(2):167–216.
- [8] Gramatika na suvremennia bulgarski knizoven ezik. 1983. *Morphologia*, tom. 2.
- [9] Stoykova V. 2002. Bulgarian noun – definite article in DATR. In Donia Scott, ed. *Artificial Intelligence: Methodology, Systems, and Applications. Lecture Notes in Artificial Intelligence 2443*, Springer-Verlag, 152–161.
- [10] Stoykova V. 2004. Modeling sound alternations of Bulgarian language in DATR. In: E. Buchberger ed. *Proceedings of KONVENS 2004, Schriftenreihe der Oesterreichischen Gesellschaft für Artificial Intelligence*, Band 5, Wien, 201-204.
- [11] Stoykova V. 2011. Interpreting Bulgarian sound alternations of inflectional morphology in DATR. In : J. Filipe and A. Fred eds. *Proceedings of the International Conference on Agents and Artificial Intelligence (ICAART)*, vol. 1, SciTePress, 486–491.
- [12] The DATR Web Pages at Sussex. 1997.  
<http://www.cogs.susx.ac.uk/lab/nlp/datr/>

**Appendix:**

```

Zakon:
<> == N_I
<infl_root> == zakon
<sem animacy> == inanimate.

Zakon: <gloss> = law.
Zakon: <mor nom sg> = zakon.
Zakon: <mor acc sg> = zakon.
Zakon: <mor gen sg> = zakon _a.
Zakon: <mor dat sg> = zakon _u.
Zakon: <mor inst sg> = zakon _om.
Zakon: <mor loc sg> = zakon _e.
Zakon: <mor nom pl> = zakon _i.
Zakon: <mor acc pl> = zakon _i.
Zakon: <mor gen pl> = zakon _ov.
Zakon: <mor dat pl> = zakon _a _m.
Zakon: <mor inst pl> = zakon _a _m'i.
Zakon: <mor loc pl> = zakon _a _x.
Zakon: <syn gender> = masc.
Zakon: <syn animacy> = inanimate.

```

```

Zakon:
<> = Noun.
<root> = zakon.

```

```

Zakon: <gender> == masc_1.
Zakon: <sing undef> == zakon.
Zakon: <plur undef> == zakon_i.
Zakon: <sing def_1> == zakon_ut.
Zakon: <sing def_2> == zakon_a.
Zakon: <plur def_1> == zakon_i_te.

```

# Recent Developments in ParaSol: Breadth for Depth and XSLT Based Web Concordancing with CWB

Ruprecht von Waldenfels

Institut für slavische Sprachen und Literaturen, Universität Bern

**Abstract.** The article describes the Slavic parallel corpus ParaSol, developed in Bern and Regensburg. The paper gives an account of recent developments, focussing on conceptual decisions concerning corpus make up and the user interface.

## 1 Introduction

ParaSol is a multilingual Slavic parallel corpus comprising original and translated prose texts developed in collaboration of the University of Bern in Switzerland and the University of Regensburg in Germany; the acronym stands for *Parallel Corpus of Slavic and Other Languages*. Having initially been developed in Regensburg under the name *Regensburg Parallel Corpus*, it is now headed in Bern University and hosted on servers in both places. Web site development and text acquisition is shared between the two institutes<sup>1</sup>.

The following principles, some of them new or modified, guide the development of ParaSol:

- *original and translated prose in many Slavic and some non-Slavic languages (breadth for depth)*
- *variation of (preferably Slavic) source languages*
- *automatic preprocessing and alignment*
- *linguistic annotation such as lemmatization and POS tagging*
- *public availability through a web concordancer*
- *crowdsourcing: users are encouraged to collaborate*

The present article focusses on text strategies and on new developments concerning the web interface.

### 1.1 Similar projects

ParaSol is most comparable to two other current projects:

---

<sup>1</sup> The following people are part of the project: Ruprecht von Waldenfels (head of the project and the Bern team; overall corpus architecture, corpus maintenance, interface design and text acquisition); Roland Meyer (head of the Regensburg team, CWB integration and interface design); Veronika Wald, Dmitrij Nikolenko (text acquisition, Regensburg); Vivian Kellenberger, Michael Reinhard, Karin Zurbuchen (text acquisition, Bern); Andreas Zeman (text acquisition and interface design, Bern).

- InterCorp [6] , developed in the Czech republic under the auspices of the Czech National Corpus, a corpus built in a cooperation of numerous departments across the country and focusing on language pairs composed of Czech and a second language, one of currently 21 European languages (see <http://www.korpus.cz/intercorp>). Alignment is done manually between these pairs (with some supplementary automatic alignment with Czech as a pivot done for texts that are present in more than one pair). Where possible, linguistic annotation (lemmatization and POS-tagging) is included. Access is provided, after registration, via a web interface.
- The Amsterdam Slavic Parallel Aligned Corpus (ASPAC), developed in Amsterdam by Adrie Barentsen. This corpus focuses on all Slavic languages and also includes some other European languages. Alignment is done manually; all language versions are aligned in a tabular fashion so as to reflect equivalence to the original (see below). As a consequence, all difficulties resulting from omissions or additions in translation, or varying source documents, are resolved explicitly. No lemmatization or POS-tagging is performed on the text. While it is not searchable through a web interface, access to the corpus is available for research on personal request (see [home.medewerker.uva.nl/a.a.barentsen/page3.html](http://home.medewerker.uva.nl/a.a.barentsen/page3.html)).

## 2 Corpus composition

### 2.1 Breadth for Depth

Like in ASPAC, and in contrast to InterCorp, both of which ParaSol cooperates closely with, the focus of the ParaSol corpus has developed to be on breadth, rather than depth, of coverage. In other words, the corpus composition strategy at this point stresses the inclusion of more language versions of a given text, rather than more texts for a given language pair. The augmentation of specific pairs of languages (as, e.g., in the past the Polish-Russian, Slovak-Bulgarian and German-Slovak pairs) was at the center of attention in the earliest phase of ParaSol (then RPC), since its rechristening as ParaSol in 2009, preference has been given to the inclusion of texts in many Slavic languages.

This development reflects a typical position of Slavic studies outside the Slavic speech communities: rather than being focussed on any of the particular national languages, our interest potentially involves all Slavic languages and, to a considerable extent, a comparative perspective on Slavic; see [8] for an approach where translation variants in diverse languages is crucial.

Moreover, our students typically study more than one Slavic language in a variety of combinations. Since ParaSol is used in pedagogical applications, especially in early stages where students do not yet have full command of these languages, having a wide range of language versions per text is an important asset. For these reasons, ParaSol has been focussing on texts that are present in many Slavic languages rather than continuing a focus on depth, that is, the addition of texts of any specific language pair.

It is therefore no coincidence that ParaSol is in this way most similar to the Amsterdam Slavic Parallel Aligned Corpus, also developed outside of the Slavic countries. Both corpora differ in this from the Czech project InterCorp, which focusses on pairs of languages with Czech.

In distinction to the projects mentioned above, ParaSol strives to balance source languages as far as practical in order to be able to deal with translation effects. As of summer

2011, 8 novels from 7 source languages<sup>2</sup> are available in translation into almost all Slavic literary languages: J.K. Rowling's *Harry Potter and the sorcerer's stone* (English), Milan Kundera's *Nesnesitelná lehkost bytí* (Czech), Mikhail Bulgakov's *Master i Margarita* (Russian), Nikolaj Ostrovskij's *Kak zakaljalas' stal* (Russian), Ivo Andrić's *Na Drini ćuprija* (Serbian); Umberto Eco's *Il nome della rosa* (Italian), Patrick Sueskind's *Das Parfüm* (German), Stanisław Lem's *Solaris* (Polish). The reader is referred to the project web sites (see below) for a current list of texts included in the corpus.

In addition to a complete coverage of each text in all (major) Slavic languages, we also strive to include German, French and Italian, the national languages of Switzerland and Germany. These are the most frequent non-Slavic native languages of our students. Moreover, we try to include Modern Greek for research interest, as this is a language especially interesting for comparison being both a member of the Balkan Sprachbund, like Macedonian and Bulgarian, and an aspect language. The Baltic languages, most closely related to Slavic, are also represented. Aside from that, we take an opportunistic stance to including other languages.

The corpus project initially grew out of the recognition that in contrastive work, researchers often compile their own small parallel corpora. ParaSol is conceived as a corpus architecture that can accommodate such projects. We continue to encourage users to contribute and make use of its facilities, adhering to a wiki spirit of crowdsourcing in corpus compilation.

### 3 Design decisions and web interface

#### 3.1 Annotation

As far as possible, texts in the corpus are lemmatized and POS tagged; where such tools are not publically available, this is done in cooperation with institutions that develop these tools in the context of the national corpora (see web site for a list of cooperations and [7] for more details).

#### 3.2 Alignment

A conceptual decision was taken to rely on pairwise alignments, rather than on a table-like alignment architecture that would involve transitive alignment properties. To assess the differences, consider the example in figure 1, where a corresponding text segment is divided into two sentences in only two of three languages.

DE	RU	PL
DE.1 Lass mich.	↔ RU.1 Pusti.	PL.1 Puść, nie chcie,
DE.2 Ich will nicht, dass Du mich berührst.	↔ RU.2 Ne choću, čtoby ty ko mne prikasalsja.	↔ zebyś mnie dotykał!

**Fig. 1.** An alignment example with differences in segmentation across languages.

<sup>2</sup> Thanks are due to Emmerih Kelih who has contributed Nikolaj Ostrovskij's *Kak zakaljalos' stal* in eleven Slavic languages [2] and a large number of translations in the *Bulgakov* subcorpus.

Let us suppose a user is interested in Russian *pusti* and Polish *puść*, cognate items both translated as *let!*. With pairwise alignment, each language version is aligned to each other language version independently. This means that if one chooses to base one's search on the Russian text, RU.1 *Pusti*. will be aligned to German DE.1 *Lass mich*. and Polish PL.1 *Puść, nie chcie, zebyś mnie dotyka!*. If, however, the search is based on Polish, the segment PL.1 *Puść, nie chcie, zebyś mnie dotyka!* is aligned to the German PL.1-2 *Lass mich. Ich will nicht, dass Du mich berührst.* and Russian RU.1-2 *Pusti. Ne choću, čtoby ty ko mne prikasalsja*. On a pairwise basis, alignment is thus maximally precise, but differs depending on which language the search is based on.

In table-like alignment, in contrast, rows such as the one in the example above are, like in a table, considered a single segment aligned across *all* versions. The more fine grained equivalence relations between Russian and German are disregarded. Therefore, any query will output the same segments regardless of which language variant the search is based on.

The decision for pairwise alignment makes the approach more robust: if any one of the language pair based alignment relations breaks down for some reason, e.g., because the text is abridged or censored, this does not result in degradation of the alignment quality of any other pair. Since ParaSol exclusively relies on automatic alignment, robustness is very important.

Alignment was initially done with bsa [3]; we have now moved to hunalign [5] (see Rosen [4] for a comparison of aligners). Where possible, alignment is done on files containing word forms replaced with lemmas in order to reduce the search space during alignment [7].

**Query interface**

Choose primary and aligned language(s), and enter a query. You need to define a query for the primary language (in red). In addition, you may define queries on the aligned languages, which will restrict output accordingly.

Primary language: **Slavonic**  BG  SRA  PLA  RU  HR  SL  SK  RUA  MK  CZ  US  UK  SR  PL  BY

**Germanic**  NL  EN  DE  DEA

**Romance**  FR  ES  IT  PT  RO

**Baltic**  LV  LT

**Others**  EO  EL  HU

Aligned languages: **Slavonic**  BG  SRA  PLA  RU  HR  SL  SK  RUA  MK  CZ  US  UK  SR  PL  BY

**Germanic**  NL  EN  DE  DEA

**Romance**  FR  ES  IT  PT  RO

**Baltic**  LV  LT

**Others**  EO  EL  HU

All texts  Only texts available in all languages

	hr	sr	sra	sl	cz	pl	us	ru	lv	el
<input checked="" type="checkbox"/> sueskindparfuem	<input checked="" type="checkbox"/>									
<input checked="" type="checkbox"/> lemglspana	<input checked="" type="checkbox"/>									
<input checked="" type="checkbox"/> bulgakovmaster	<input checked="" type="checkbox"/>									
<input checked="" type="checkbox"/> ostrovskijstal	<input checked="" type="checkbox"/>									
<input checked="" type="checkbox"/> pavichazar	<input checked="" type="checkbox"/>									
<input checked="" type="checkbox"/> lemfiasko	<input checked="" type="checkbox"/>									
<input checked="" type="checkbox"/> lemsolaris	<input checked="" type="checkbox"/>									
<input checked="" type="checkbox"/> potter1	<input checked="" type="checkbox"/>									

Serbian

Croatian

Serbian a

Slovene

Czech

Polish

Upper Sorbian

Latvian

Greek

Russian

Search

Fig. 2. Query for *[N]ikad.\** in Serbian, with a variety of aligned languages, not all present in all texts.

### 3.3 Query interface

The design of the interface<sup>3</sup> reflects the conceptual decision for pairwise alignment. The user first chooses some primary language, and then selects a set of aligned languages. As the user selects and deselects languages, the list of corpus files on the lower left side of the interface and input fields for the query strings on the lower right side appear and adapt to reflect the user's choices. This is implemented in javascript and partly backed by entries in an SQL-data base. There is an option to either restrict the set of texts to those texts that are available in all languages, or to perform the query on all texts which are present in the primary language (see the screenshot in figure 2).

Input fields for the query strings accept standard CQP syntax and directly channels queries to CWB[1], which now fully supports unicode encoded corpora. Annotation varies from language to language, but typically, three levels are supported: word form, lemma, and morphosyntactic tag. Each query opens a new result window (a feature inspired by the RNC).

### 3.4 XML/XSLT based concordance

The interface, originally developed essentially as a wrapper for the HTML output module of CWB, now utilizes client-based XSLT for the display of the XML encoded result returned by CWB. As of the moment of writing, however, CWB does not yet support XML output (although this is a planned feature, Stefan Evert, p.c.). Instead, the SGML output module is used, which, however, is faulty in respect to entity resolution. Regular expressions in the php code are used to derive valid XML from this faulty SGML representation. The resulting XML text is transferred to the client together with an XSLT style sheet that transforms it to HTML.

While the transformation from SGML to XML slows down output considerably, the transition to an XML based output system is justified by a number of advantages. First, this decision amounts to dividing content generation (the XML file) from output display (the HTML file resulting from the XSLT transformation), thus adding to the modularity of the system. The question of output generation is for the time being solved in a provisional way with transformations from SGML; this will have to be reviewed as soon as an XML module is ready. Since content generation and display are separate issues now, this temporary solution does not stand in the way of further development of the display module. Also, since XSLT is a language without side effects, directly geared to manipulating structured data, using a XSLT style sheet is much simpler, and at the same time more flexible and more robust than php code.

As an example, consider queries where not all texts are available in all languages the user is interested in, as in the query of the screen shot in figure 2. In order to format the resulting table, a server based php solution would have to keep track of which corpus is available in which language and check for consistency with the actual result table returned by CWB - for a variety of reasons, this can fail, and strategies to deal with this have to be employed. In contrast, a client side XSLT solution works locally on the resulting XML file alone. As long as this is a valid XML file, all necessary display decisions such as widths of the columns or the column labels can be taken on the data alone; since this involves

<sup>3</sup> The web interface has been developed by Roland Meyer, Regensburg; Andreas Zeman, Bern; Ruprecht von Waldenfels, Bern

much less assumptions and variables, this much more robust and at the same time easier to implement.

As a whole, moving to XML and XSLT technology has in our case resulted in much more rapid and flexible evolution of the concordance window (as shown in the screenshot in figure 2). Lemmas and morphological tags are now shown as tool tips, and basic statistics are computed on the basis of the result file on the client side.

The style sheet is much simpler and, owing to the fact that XSLT has no side effects, more robust than a server side construction of a HTML file. Moreover, since this is a modular solution, we can very easily offer more output formats now by simply adding an option to use different style sheets, which may ultimately even may be user developed or user supplied.

80159 Ni u kom slučaju i nikada !	Ни у каком случаю и никогда !	Nekad un nekādā gadī jumā !	V nobenem primeru in nikoli !	Ni u kom slučaju i nikada !	Nigdy , w żadnym wypadku !	- Už nikdy , v žádném případě nevezmu ve vašem bufetu nic do úst !	Ni u kom slučaju i nikada !	Σέ καμιά , περίπτωσή και no té !
80203 Sir ni u kom slučaju i nikada ne može da bude zelene boje , to vas je neko prevario .	Брынза не бивае зеленогo цвета , это вас кто - то обманул .	Brinza! taču nav jābūt zaļš krāsā . Jūs esat maldināts .	Ovčji sir ni zelene barve , nekdo vas je moral potegniti .	Ne postoji ovčiji vas je to neko prevario .	Zielona bryndza nie istnieje , ktoś musiał pana oszukać .	Můj milý , bryndza nesmí být nazelenaná , to vás někdo ošidil .	Ovčji sir nije zelene boje , to vas je netko obmanuo .	Δέν υπάρχει φέτο πράσινου χρώματος , κάποιος σός ξεγέλασε .
<b>40 hits in corpus <i>lensolaris</i>.</b>								
	ru	pl	cz	hr	sra			
1899 Nikad pre ga nisam video .	Раньше я никогда не видел Снаута .	Nigdy go jeszcze nie widział   em .	Osobně jsem se s ním ještě nikdy nesetkal .	Nikada ga još nisam vidio .	Nikad ga još nisam bio video .			
9867 Ovaj aneks prvog solarističkog godišnjaka bio mi je poznat , to jest znao sam za njegovo postojanje , ali ga nikad nisam imao u rukama , pošto je predstavljao čisto istorijsku vrednost .	О приложении к первому тому " Соларистического ежегодника " я знал , то есть слышал , что оно существует , но никогда не держал его в руках , поскольку оно представляло собой только историческую ценность .	Ów aneks do pierwszego solary - styczniego rocznika znał   em , to znaczy , wiedział   em o jego istnieniu , ale nie miał   em   miałem go nigdy w ręce , przedstawiał bowiem czysto historyczną wartość .	Onen dodatek k prvnímu ročníku Solaristické ročenky jsem znal , či spíše věděl jsem o jeho existenci , ale neměl jsem jej nikdy v rukou , měl totiž už jen historickou cenu .	Taj sam dodatak prvome solarističkom godišnjaku poznao , to jest , znao sam za njegovo postojanje , ali ga nikada nisam imao u ruci , jer je predstavljao čisto povijesnu vrijednost .	Ovaj aneks uz prvi solaristički godišnjak znao sam , to jest bilo mi je poznato da postoji , ali nikad ga nisam imao u ruci , jer je predstavljao čisto istorijsku vrednost .			
9891 A za nekog pak Ravincera , niti za njegov Mali apokrif nikad nisam ni čuo .	Однако я понятия не имел ни о Равинцере , ни о " Малом Апокрифе " .	Natomiast o jakimś Ravintzerze ani o jego , Małym Apokryfie " nigdy nawet nie słyszał   em .	Zato však o žádném Ravintzerovi ani o jeho " Malém apokryfu " jsem v životě neslyšel .	Međutim , za nekakvoga Ravintzera i njegov Mali apokrif , nisam nikada čak ni čuo .	Međutim o nekome Ravinceru i o njegovom " Malom apokrifu " nikad nisam bio čak ni čuo .			
12592 Nikad nisam bio na Stanici , ali sam šest nedelja stanovao u njenoj vernoj kopiji koja se nalazila u Institutu , na Zemlji .	Я никогда не был на Станции , но во время подготовки прожил шесть недель в ее точной копии , находящейся в Институте	Nie był   em nigdy na Stacji , ale przez sześć tygodni mieszkał   em — w ramach wstępnego treningu — w jej dokładnej kopii ,	Nebyl jsem nikdy na stanici , ale při včivním jsem bydlil šest týdnů v její přesné kopii , kterou mají v Institutu na Zemi .	Nikada nisam bio na Postaji , ali sam šest mjeseci - u okviru pripremnih vježbi - boravio u njezinoj točnoj kopiji , koja se nalazila u	Nisam nikad bio na Stanici , ali šest sedmica sam proveo - u okviru uvodne pripreme - u njenoj tačnoj kopiji koja se nalazila u			

Fig. 3. Query result for *[Nn]ikad.\** in Serbian, with differing number of aligned languages.

## 4 Summary

The present article has given a short overview of the ParaSol, a Parallel Corpus of Slavic Languages, focussing on two recent developments: a change in the corpus composition strategy with an aim to include more language versions of a given text, rather than more texts for a given language pair (breadth for depth) as well as a move to XML/XSLT technology for the web concordancer.

## References

- [1] Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. In *COMPLEX'94*.
- [2] Kelih, E. (2009). Slawisches parallel-textkorpus: Projektvorstellung von „kak zakaljalas' stal' (kzs)“. In Kelih, E., Levickij, V., and Altmann, G., editors, *Metody analizu teksta/Methods of Text Analysis*, pages 106–124, Černivci. ČNU.
- [3] Moore, R. (2002). Fast and accurate sentence alignment of bilingual corpora. *Machine Translation: From Research to Real Users*, pages 135–144.
- [4] Rosen, A. (2005). In search of the best method for sentence alignment in parallel texts. In Garabík, R., editor, *Computer Treatment of Slavic and East European Languages: Third International Seminar, Bratislava 10–12 November 2005*, pages 174–185, Bratislava.
- [5] Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., and Trón, V. (2005). Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing IV: Selected Papers from RANLP 2005*, pages 590–596.
- [6] Vavřín, M. and Rosen, A. (2008). Intercorp: A multilingual parallel corpus project. In *Proceedings of the International Conference Corpus Linguistics – 2008*, pages 97–104. St. Petersburg State University.
- [7] von Waldenfels, R. (2006). Compiling a parallel corpus of slavic languages. text strategies, tools and the question of lemmatization in alignment. In Brehmer, B., Ždanova, V., and Zimny, R., editors, *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV) 9*, pages 123–138. München.
- [8] von Waldenfels, R. (t.a.). Aspect in the imperative across slavic – a corpus driven pilot study. *Oslo Studies in Language. Special Issue ed. by A. Grønn & D. Haug*.

# The Instrumental Environment for the Automatic Syntactical Analysis of Ukrainian

Iryna Zamaruieva and Olga Shypnivska

National Taras Shevchenko University of Kyiv

**Abstract.** In the paper the instrumental environment for the automatic syntactical analysis of Ukrainian is described. The description of main characteristics of databases for the automatic syntactical analysis and general principles of automatic syntactical analysis are presented. Just now the group of the machine translation system of military area is preparing the pilot linguistic databases.

## 1 The automatic syntactical analysis. Some problems and proposed solutions

The question of the automatic syntactical analysis (annotation) has been a key problem for natural language processing for several years [1; 2; 3]. Automatic syntactical analysis is a useful tool, especially with regard to a corpus data-processing. In this respect, syntactical annotation has also been considered in the course of the development of the knowledge-based machine translation systems. The theoretical aspects of syntactical analysis and its application in the natural language processing were presented in many works and applied to many systems [2; 3; 4; 5; 8]. Real system of syntactical analysis for languages with a free word order is rare.

Generally syntactical analysis is understood as a formal description of a particular language. The main tasks of the syntactical analysis are to build a formal syntactical structure of a text and to represent sentences in the syntactical categories. In our work we consider automatic syntactical analysis as an independent module natural language processing (NLP) systems which can identify and generate all correct syntactical structures of Ukrainian.

The creation of the automatic syntactical analysis module deals with many unsolved theoretical and practical problems in applied linguistics. Current researches in Slavic, reported in the literature, provide descriptions for the automatic syntactical analysis [3; 4]. There are a lot of works which represent attitudes and methods of its implementation. Dependency grammar and immediate constituents grammar are used as theoretical basis of these studies [3; 4].

Ukrainian as a free word order language has many types of linear structure of sentences both simple and complete. There are a lot of structure components and their combinations in the texts. This high level of syntactical complexity is difficult to anticipate and simulate. The application of the automatic syntactical analysis indicates that many structures such as complex predicates, complex subjects, coordinative

structures, many cases of syntactical ambiguity requires special technique for preview determining and modelling.

That's why, a description of a formal syntactical structure of Ukrainian is impossible without good special tools. As a suitable variant we present the instrumental environment. In the paper we offer a description of the morphological tagging, general principles of automatic syntactical analysis according to the knowledge-based machine translation system. We propose a number of linguistic databases for automatic syntactical analysis as a research environment and demonstrate their advantages. The textual data used for the task consists of military texts with the morphological tagging. Balabin V. V., Zamarueva I. V., Tolubko V. V., Ljashenko A. V., Shypnivska O. O. are involved to the project.

## 2 Description of morphological parser and morphological tagging of Ukrainian

There are several requirements which are necessary for effective functioning of morphological parser, including multitopic dictionary for analysis texts of different types, the system have to be open and inflection, new words must be parsed. Never the less the choice of the morphological parser depends on practical reasons. For the development of machine translation system we prefer inflectional attitude with the kwasiflexes dictionary. Offered method is realizing due to the positional-numeral encoding of the grammatical information in the dictionary entry. Grammar information encoding is common for Ukraine, English, Russian. Part-of-speech codes and their immanent categories displays table 1. Symbol "+" marks positions which grammar category meaning complites. In the code this symbol is assigned number from 1 till 9. In case when part-of-speech doesn't have certain category the position is assigned zero.

Part-of-speech code	Part-of-speech	Grammar category									
		Gender	Number	Case	Person	Mood	Voice	Degree	Tense	Reflexivity	
1	Noun	+	+	+							
2	Adjective	+	+	+							
3	Numeral-noun	+		+							
4	Numeral- Adjective	+	+	+							
5	Pronoun	+	+	+	+						
6	Pronoun-Noun			+							
7	Pronoun- Adjective	+	+	+							

8	Verb in the past	+	+			+				+	
9	Verb not in the past		+		+	+	+		+	+	
10	Infinitive					+					
11	Verb in imperative mood		+								
12	Participle	+	+	+			+		+		
13	Adverbial participle										+
14	Adverb										
15	Adjective in comparative degree							+			
16	Possessive Adjective	+	+	+							
17	The short form of the adjective	+	+								
18	The short form of the participle	+	+				+			+	
19	Adverb in comparative degree							+			
20	Particle										
21	Modal word										
22	Article										+

**Table 1.** Part-of-speech codes for the morphological parser

Kwasiflexes dictionary was formed on the base of word forms dictionary. For the creation of this resource the stylistic differentiated texts were parsed manually and inverse word forms dictionary was generated. Every token is getting his own positional-numeral code which contains grammar information about its part-of-speech and particular meaning. In case of the morphological ambiguity there are several codes. In code on the first position the part-of-speech is indicated. Than every particular grammar meaning is denoted by numeral. Kwasiflexes dictionary was formed from this dictionary. Offered inflectional attitude allows us:

- to shorten analytical dictionary volume and expedite morphological parser operation;
- to create an open system which is able to analyze new words.

In this way Ukrainian kwasiflexes dictionary which contains 5.5 thousand of kwasiflexes was performed. Picture 1 displays a fragment of kwasiflexes dictionary. Besides morphological parser has word-formative patterns dictionary and dictionary of syntactic words.

аб	1*21/ба/1*11/б/2*21/бий/ба/бе/1*41/ба/
аберго	1*11/огрїб/
абер	1*11/реб/
або	1*11/їб/1*41/оба/1*21/оба/
абол	1*11/лоб/1*21/лоба/
абор	1*41/роба/1*11/рїб/
аборк	1*11/кроб/
аборо	1*11/ороб/1*21/ороба/
абоф	1*11/фоб/

**Picture 1.** Ukrainian kwasiflexes dictionary

### 3 General description of the automatic syntactical analysis and some result of its implementation

According to our approach of performing machine translation system automatic syntactical analysis consists of three stages:

- determining syntactical connections between word forms in a sentence – contextual analysis;
- building a formal syntactical structure of sentences;
- building a formal structure of complex syntactical units.

Contextual analysis is the first step towards the first level of our bases. It is proceed fully automatically. For every couple of words a type of syntactical relations and a governing word are determined. Morphological ambiguity is resolved at this step of analyzing.

The identification of the syntactical attitudes is realized due to the usage of syntactical rules dictionary which contains contextual-syntactical rules of agreement, government, parataxis. The format of such conditions representation is shown in the table 2. The first column provides part-of-speech code of a word, the second column provides part-of-speech code of a word due to which the syntactical analyze is implemented and the third one shows part-of-speech code of a word with which syntactical attitude is realized. The forth column provides grammar meaning according to which syntactical attitude is realized. In the fifth and sixth columns the type of syntactical attitude and governing word are contained. The seventh column shows the type of operation. The declarative method allows us to do this step of syntactical analysis using only the table.

The picture 2 displays the implementation of the syntactical rules dictionary usage.

Part-of-speech	Due to which part-of-speech	With what part-of-speech	According to which grammar information			Type of syntactical attitude	Governing word	Operation
			Gender	Number	Case			
1*		2*	+	+	+	У	1*	M1
1*	24*	1*			+	У	1*/2	M1
...		...	...	...	...	...	...	...
1*		23*			2	К	23*	M1
1*		1*			2	К	1*/2	M2
...		...	...	...	...	...	...	...
14*		9*				П	9*	M3

**Table 2.** The format of the representation of contextual connection rules

```

A =>
C1 Фактичне 2*311000000\2*314000000\ (ГС) командування
У1 (ГС) командування
1*311000000\1*313000000\1*314000000\1*321000000\1*324000000\ видами
1*125000000\
У1 (ГС) видами 1*125000000\ ЗС 41*
У1 (ГС) ЗС 41* США 40*
належить 9*010329012\
У1 (ГС) начальникам 1*123000000\ штабів 1*122000000\
, МС => які входять 9*020329012\
У2 (ГС) до 23*002000000\1*399000000\ комітету 1*112000000\1*113000000\
С1 об'єднаного 12*112002030\12*114002030\12*312002030\ (ГС) комітету
1*112000000\1*113000000\
У1 (ГС) комітету 1*112000000\1*113000000\ начальників
1*122000000\1*124000000\
У1 (ГС) начальників 1*122000000\1*124000000\ штабів 1*122000000\
(ОКНШ) 41*
- МС => С1 консультативного 2*112000000\2*114000000\2*312000000\ (ГС)
органу 1*112000000\1*113000000\
У1 (ГС) органу 1*112000000\1*113000000\ президента
1*112000000\1*114000000\
У2 (ГС) з 23*02000000\23*04000000\23*06000000\ питань 1*322000000\
С1 військових 2*922000000\1*922000000\ (ГС) питань 1*322000000\...[KP]
    
```

**Picture 2.**

The results of the contextual analysis are processed by a module of an interpretation. Word-combinations, terms and concepts of the certain object region are determined at this step. For this reason every word of syntactical constructions is converted to its lemma and correlated with the semantic interpretation dictionary. Identified syntactical constructions are interpreted as one lexeme and grammar description is imputed to a governing word. Word-combinations which contains verbs is comparing with the valency dictionary. The input of the interpretation dictionary is

grammar information and output is semantic information. The result of the contextual-syntactical analysis is shown in the picture 3.

```

MP =>  MC =>
(У) Важливий[2*221000000/2*224000000/] (ГС)функція[1*224000000/1*221000000/]
(К) (ГС)у[23*006000000/] (ГС)забезпечення[1*316000000/]
національна безпека[5/1*212000000/]
(П) (ГС)виконувати[1/9*024329012/9*026329012/9*022329012/] також[14*000000000/]
Конституційний[69/2*111000000/69*114000000/]
Суд[69/1*111000000/69*114000000/]
Україна [62/1*232000000/]
.[L16]->  MC =>
(К) Прокуратура[69/1*211000000/] Україна[62/1*232000000/]
.[L16]->      MC =>
(У) (ГС)Національний[69/2*111000000/1*114000000/]
(К) банк[1*111000000/1*114000000/] Україна[62/1*232000000/]
.[L16]->      MC =>
(У) (ГС)міністерство[1*312000000/1*321000000/1*324000000/] і[24*000000000/]
відомство[1*312000000/1*321000000/1*324000000/] .[КР]

```

**Picture 3.** The result of the contextual-syntactical analysis

The sentence from this picture contains four word-combinations which are considered as one lexeme.

On the second step of the automatic syntactical analysis the syntactical structure of simple sentence is built. The predicate, subject, object, attribute, circumstance are determined. This step was applied to only simple sentences.

The application of these two steps displays us there are many syntactical structures both formal and phrasal which require special previous analysis. In many cases it's necessary to build separate modules including different linguistic data. That is why we decided to create linguistic databases for automatic syntactical analysis as a responsible tool for researching.

#### **4 Forming the databases for the automatic syntactical analysis of Ukrainian**

We consider the linguistic databases for automatic syntactical analysis as a research environment. These linguistic databases generally designed:

- to represent a list of linear structure patterns of Ukrainian sentences – simple, complete;
- to contain all right syntagmas which are necessary both for analysis and synthesis of Ukrainian;

- to show all possible cases of syntactical ambiguity and their occurrence probability in the real texts;
- to possess all syntactical structures for which it is necessary to form special module of their processing.

These databases were formed both manually and automatically. They consist of 4 tables. The first table contains simple sentences, the second one – compound sentence, the third – complex sentence and the fourth – complicated sentence. The textual data used for the task consists of military texts. The input of our bases is sentences after the morphological analysis. Manually for every sentence its type was determined and added. Criteria of classification you can see on the table 3. Based on previous studies we assume that general criteria can be for us as a base for the generalization.

Criteria of classification	Type of a sentence
Number of predicates	Simple
	Complex
	Another
With/without auxiliary components	Extended
	Unextended
	Another
Type of narration	Narrative
	Interrogative
	Illocutionary
	Another
Affective evaluation	Exclamatory
	Unexclamatory
	Another

**Table 3.**

We try to do universal format of data presentations for all types of sentences and consider simple sentences as a basic syntactical structure relative to complete sentences. As we said determination a type of a sentence is obligatory for all tables. Besides, for simple sentences we determine their character according availability/unavailability of predicate or subject, for complete sentences we determine a type of clause, a type of a syntactical relationship (conjunctive/unconjunctive) is determined. The format of data presentations for simple sentences you can see on the table 4.

Sentence
Type of sentence according to narration
Type of sentence according to availability/unavailability of predicate or subject
Type of predicate
Type of subject
Position of predicate according to subject
Pattern of coordinative construction
Type of syntactical ambiguity
Punctuation

**Table 4.**

Linguistic data organized in this way allows us to obtain any information as for syntactical structure, considering their occurrence probability. We also can prove or disprove some theoretical, practical conclusions or obtain new conclusions. For example, we can take only simple sentences. We distinguish two kinds of simple sentences according availability/unavailability of predicate or subject. Sentences which have both predicate and subject present 90% of all data. Among them more than 53% coverage sentences which have secondary parts of sentences and are complicated by participle and adverbial participle phrases. This fact let us rethink theoretical reasoning of simple sentences once more. For example, we picked out 16 patterns of coordinative construction with secondary parts of sentences for which it is necessary to form special module of their processing [6].

We attach particular attention to main parts of sentences – predicate and subject. Finding of these parts specifies correct understanding of a sentence and finally a text in general. It's important to determine patterns of predicate and subject, manner of their coordination and position as to each other [7]. Undertaking into account one of this fact can substantially change information. For example as we can see in the sentence taken after the machine translation system from Russian into Ukrainian. In Russian we have: “Вот как это трагическое событие описывает летопись Китая: “Переместились горы и реки, дороги были разрушены”. In Ukrainian the subject “летопись” became the object but the object “событие” became the subject: “От як це трагічна подія описує літопис Китаю: “Перемістилися гори і річки, дороги були зруйновані.”

In our databases we identified the peculiarity of predicate occurrence probability which is represented in the table 5. At this moment these databases contain 100 000 word forms and have above 500 sentences.

Type of predicate	Type of subject	Occurrence probability	Examples
Simple verbal	Simple nominal	71%	дії розпочинаються
Complex nominal	Simple nominal	22%	прикладом є громадянська війна
Complex verbal	Simple nominal	5.4%	сторони зобов'язані дотримуватися
Simple verbal	Complex nominal	1.4%	207 гелікоптерів будуть оснашені
Complex nominal	Complex nominal		

Table 5.

From this table we can see that simple predicate and subject coverage 56% of all patterns. Sentences with the predicate in the second position as to the subject represent 68% of all occurrences. It allows us to disprove opinion that the predicate as a rule is used in the first position in Slavic. We have to remember that 32% where the predicate keeps the first position tell us to mark sentences in some cases as ambiguous. Especially it's necessary when subject and object are used in the same form – in nominative. Let you see next sentences: Ствол визначає напря́м польоту кулі. Шикуння́ підрозділів на марші забезпечує можливість вступати у бій з ходу. У відповідності з законом України “Про оборону України” та “Про Збройні Сили України” загальне керівництво Збройними силами України здійснює Президент як Головнокомандувач Збройних Сил України.

Both analysis and synthesis texts in Ukrainian need to create all possible syntactical structures. In this case there is a necessity to consider syntactical ambiguity. For discovering of ambiguity constructions we propose our databases in which special fields are assigned. First of all we have to admit that syntactical ambiguity is rather complicated theoretical questions. In our work we investigate it concerning to practical requirements. That it's why we differ two types of ambiguity: type 1 – constructions where it is difficult to determine a governing word in constructions or in sentences; type 2 – constructions where it is difficult to determine a type of syntactical connections between word forms in a constructions or in sentences. Let us compare in the first sentence we have to know that the word “поранення” is independent: “Радіоактивні речовини проникають у організм через органи дихання, травлення та поранення” and does not depend from the word “органи”. In the second construction: “сорочку кольору полину з погонами і галстуком кольору полину без парадно-вихідної куртки при літній парадно-вихідній формі одягу” we have to define that the combination “погонами” depends from “сорочку” but not from “полину”.

It is important to recognize the depth of the syntactical analysis particularly in constructions with a preposition like that: “Вася п’є каву з молоком, Вася п’є каву з тістечком, Вася п’є каву з Тетяною, Вася п’є каву із задоволенням”. In these cases constructions with a preposition can represent different syntactical connections: attributive, objective, subject, adverbial modifier. As a rule native speaker does not note this peculiarity of language but for automatic syntactical analysis it’s rather complex problem. Let see next sentences with such constructions. In first one “Ударно–спусковий механізм слугує для спуску *курка з бойового взводу*” we distinguish for the word “курка” two types of relations with the next construction: attributive and adverbial. In the second case: *пересування на лижах із невеликою швидкістю* – we have to know the governing word for the construction “*із невеликою швидкістю*” there is the word “*пересування*” or the word “*лижах*”. We have also to determining the type of attitude.

In our work this considered characteristic we apply relative to complete sentences in tables 2, 3, 4. For example we take the table with complete sentences with ambiguity. In the first sentence “Оскільки танки Т-84 та Т-80У мають ряд спільних технічних рішень, доцільно розглянути їх разом” it is difficult to identify a governing word for the subordinate clause “доцільно розглянути їх разом”. We have two variants of the governing word: “танки” and “рішень”. In the second one “Суккупність органів державної влади, військових формувань, утворених відповідно до законів України, діяльність яких перебуває під демократичним цивільним контролем з боку суспільства і безпосередньо спрямована на захист національних інтересів України від зовнішніх загроз, називається військовою організацією держави” there are two questions; we have to specify a governing word for the participle phrase “утворених відповідно до законів України” (“органів” or “формувань”) and we have to specify a governing word for a subordinate clause “діяльність яких перебуває...”. In both cases sentences are marked as ambiguous and will be analyzed later.

A linear structure pattern of a sentence we understand as a linear succession of secondary parts of sentence and phrases: adverbial participle phrases or participle phrases if they are. For this time we realized this task of our databases only for simple sentences. Next pattern represents a simple sentence with a participle phrase “Основні засади державної політики, спрямованої на захист національних інтересів і гарантування в Україні безпеки особи, суспільства і держави від зовнішніх і внутрішніх загроз, визначаються Законом України “Про основи національної безпеки України” від 19 червня 2003 року” Atr Subj<sub>1</sub> Atr AtrF Predic<sub>1</sub> Obj Atr (where Subj<sub>1</sub> – simple nominal subject, Predic<sub>1</sub> – simple verbal predicate, AtrF – a participle phrase).

## 5 Further developments

Our instrumental environment can obviously be used for further linguistic research. Besides, these linguistic databases can be integrated to different current natural language processing systems. Using the result of this preliminary analysis of a sentence a new table of syntactic information was designed. Every complex sentence in this data is presented by its own models as combination linear periods. Every period, punctuation characters as a syntactical indicators have their own field. This attitude allows us describe different separated parts of sentences. We have effectively been forced to show the syntactic behavior of Ukrainian more explicitly and more widely (including peripheral phenomena) than ever. After these studies syntactical functions of every word in the sentence such as predicate, subject, complement, attribute, adverbs are determined. These databases are considered to be an auxiliary, intermediate step towards the deep syntactic level.

## References

- [1] J. D. Apresan, I. M. Boguslavskij (1978), *Lingvisticheskoje obespečenije w sisteme awtomaticheskogo perevoda tretjego pokolenija*, Moscow.
- [2] A. V. Gladkij (1985), *Sintaxsicheskije structury jestestvennogo jazyka w awtomatizirovannyh sistemach obszchenija*, Moscow.
- [3] *Sintaxsicheskij analiz nauchnogo texta na EBM* (1999), Kyjiv.
- [4] A. M. Batalina, M. E. Epiphonov, T. J. Kobzareva, E. V. Kushnareva, D. G. Lachuti (2006), *Opyt experimentalnoj realizaciji algoritmov poverchnostno-sintaxsicheskogo analiza* [in:] [www.dialog-21.ru/Archive/2006](http://www.dialog-21.ru/Archive/2006)
- [5] T. J. Kobzareva (2007), *Ijererchija zadach poverchnostno-sintaxsicheskogo analiza russkogo predlozenija* // *NTI. Ser. 2., № 1.*
- [6] V. Z. Sannikov (1989), *Russkije sochinitelnyje konstrukciji. Semantica. Pragmatica. Sintaksis.*, Moscow.
- [7] L. L. Iomdin, I. A. Melchuk, N. B. Percov (1975), *Fragment modeli russkogo poverchnostnogo syntaxsisa. Predikativnyje syntagmy* [in:] *NTI. Ser. 2., № 7.*
- [8] J. Hajič (2006), *Complex Corpus Annotation: The Prague Dependency Treebank* [in:] *Insight into the Slovak and Czech corpus linguistics*, Bratislava.

## Appendix



### Attribution-ShareAlike 3.0 Unported

CREATIVE COMMONS CORPORATION IS NOT A LAW FIRM AND DOES NOT PROVIDE LEGAL SERVICES. DISTRIBUTION OF THIS LICENSE DOES NOT CREATE AN ATTORNEY-CLIENT RELATIONSHIP. CREATIVE COMMONS PROVIDES THIS INFORMATION ON AN “AS-IS” BASIS. CREATIVE COMMONS MAKES NO WARRANTIES REGARDING THE INFORMATION PROVIDED, AND DISCLAIMS LIABILITY FOR DAMAGES RESULTING FROM ITS USE.

### License

THE WORK (AS DEFINED BELOW) IS PROVIDED UNDER THE TERMS OF THIS CREATIVE COMMONS PUBLIC LICENSE (“CCPL” OR “LICENSE”). THE WORK IS PROTECTED BY COPYRIGHT AND/OR OTHER APPLICABLE LAW. ANY USE OF THE WORK OTHER THAN AS AUTHORIZED UNDER THIS LICENSE OR COPYRIGHT LAW IS PROHIBITED.

BY EXERCISING ANY RIGHTS TO THE WORK PROVIDED HERE, YOU ACCEPT AND AGREE TO BE BOUND BY THE TERMS OF THIS LICENSE. TO THE EXTENT THIS LICENSE MAY BE CONSIDERED TO BE A CONTRACT, THE LICENSOR GRANTS YOU THE RIGHTS CONTAINED HERE IN CONSIDERATION OF YOUR ACCEPTANCE OF SUCH TERMS AND CONDITIONS.

#### 1. Definitions

- a. **“Adaptation”** means a work based upon the Work, or upon the Work and other pre-existing works, such as a translation, adaptation, derivative work, arrangement of music or other alterations of a literary or artistic work, or phonogram or performance and includes cinematographic adaptations or any other form in which the Work may be recast, transformed, or adapted including in any form recognizably derived from the original, except that a work that constitutes a Collection will not be considered an Adaptation for the purpose of this License. For the avoidance of doubt, where the Work is a musical work, performance or phonogram, the synchronization of the Work in timed-relation with a moving image (“synching”) will be considered an Adaptation for the purpose of this License.
- b. **“Collection”** means a collection of literary or artistic works, such as encyclopedias and anthologies, or performances, phonograms or broadcasts, or other works or subject matter other than works listed in Section 1(f) below, which, by reason of the selection and arrangement of their contents, constitute intellectual creations, in which the Work is included in its entirety in unmodified form along with one or more other contributions, each constituting separate and independent works in themselves, which together are assembled into a collective whole. A work that constitutes a Collection will not be considered an Adaptation (as defined below) for the purposes of this License.
- c. **“Creative Commons Compatible License”** means a license that is listed at <http://creativecommons.org/compatiblelicenses> that has been approved by Creative Commons as being essentially equivalent to this License, including, at a minimum, because that license: (i) contains terms that have the same purpose, meaning and effect as the License Elements of this License; and, (ii) explicitly permits the relicensing of adaptations of works made available under that license under this License or a Creative Commons jurisdiction license with the same License Elements as this License.
- d. **“Distribute”** means to make available to the public the original and copies of the Work or Adaptation, as appropriate, through sale or other transfer of ownership.

e. **“License Elements”** means the following high-level license attributes as selected by Licensor and indicated in the title of this License: Attribution, ShareAlike.

f. **“Licensor”** means the individual, individuals, entity or entities that offer(s) the Work under the terms of this License.

g. **“Original Author”** means, in the case of a literary or artistic work, the individual, individuals, entity or entities who created the Work or if no individual or entity can be identified, the publisher; and in addition (i) in the case of a performance the actors, singers, musicians, dancers, and other persons who act, sing, deliver, declaim, play in, interpret or otherwise perform literary or artistic works or expressions of folklore; (ii) in the case of a phonogram the producer being the person or legal entity who first fixes the sounds of a performance or other sounds; and, (iii) in the case of broadcasts, the organization that transmits the broadcast.

h. **“Work”** means the literary and/or artistic work offered under the terms of this License including without limitation any production in the literary, scientific and artistic domain, whatever may be the mode or form of its expression including digital form, such as a book, pamphlet and other writing; a lecture, address, sermon or other work of the same nature; a dramatic or dramatico-musical work; a choreographic work or entertainment in dumb show; a musical composition with or without words; a cinematographic work to which are assimilated works expressed by a process analogous to cinematography; a work of drawing, painting, architecture, sculpture, engraving or lithography; a photographic work to which are assimilated works expressed by a process analogous to photography; a work of applied art; an illustration, map, plan, sketch or three-dimensional work relative to geography, topography, architecture or science; a performance; a broadcast; a phonogram; a compilation of data to the extent it is protected as a copyrightable work; or a work performed by a variety or circus performer to the extent it is not otherwise considered a literary or artistic work.

i. **“You”** means an individual or entity exercising rights under this License who has not previously violated the terms of this License with respect to the Work, or who has received express permission from the Licensor to exercise rights under this License despite a previous violation.

j. **“Publicly Perform”** means to perform public recitations of the Work and to communicate to the public those public recitations, by any means or process, including by wire or wireless means or public digital performances; to make available to the public Works in such a way that members of the public may access these Works from a place and at a place individually chosen by them; to perform the Work to the public by any means or process and the communication to the public of the performances of the Work, including by public digital performance; to broadcast and rebroadcast the Work by any means including signs, sounds or images.

k. **“Reproduce”** means to make copies of the Work by any means including without limitation by sound or visual recordings and the right of fixation and reproducing fixations of the Work, including storage of a protected performance or phonogram in digital form or other electronic medium.

**2. Fair Dealing Rights.** Nothing in this License is intended to reduce, limit, or restrict any uses free from copyright or rights arising from limitations or exceptions that are provided for in connection with the copyright protection under copyright law or other applicable laws.

**3. License Grant.** Subject to the terms and conditions of this License, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) license to exercise the rights in the Work as stated below:

- a. to Reproduce the Work, to incorporate the Work into one or more Collections, and to Reproduce the Work as incorporated in the Collections;
- b. to create and Reproduce Adaptations provided that any such Adaptation, including any translation in any medium, takes reasonable steps to clearly label, demarcate or otherwise identify that changes were made to the original Work. For example, a translation could be marked “The original work was translated from English to Spanish,” or a modification could indicate “The original work has been modified.”;
- c. to Distribute and Publicly Perform the Work including as incorporated in Collections; and,
- d. to Distribute and Publicly Perform Adaptations.
- e. For the avoidance of doubt:

1. **Non-waivable Compulsory License Schemes.** In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme

cannot be waived, the Licensor reserves the exclusive right to collect such royalties for any exercise by You of the rights granted under this License;

ii. **Waivable Compulsory License Schemes.** In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme can be waived, the Licensor waives the exclusive right to collect such royalties for any exercise by You of the rights granted under this License; and,

iii. **Voluntary License Schemes.** The Licensor waives the right to collect royalties, whether individually or, in the event that the Licensor is a member of a collecting society that administers voluntary licensing schemes, via that society, from any exercise by You of the rights granted under this License.

The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. Subject to Section 8(f), all rights not expressly granted by Licensor are hereby reserved.

**4. Restrictions.** The license granted in Section 3 above is expressly made subject to and limited by the following restrictions:

a. You may Distribute or Publicly Perform the Work only under the terms of this License. You must include a copy of, or the Uniform Resource Identifier (URI) for, this License with every copy of the Work You Distribute or Publicly Perform. You may not offer or impose any terms on the Work that restrict the terms of this License or the ability of the recipient of the Work to exercise the rights granted to that recipient under the terms of the License. You may not sublicense the Work. You must keep intact all notices that refer to this License and to the disclaimer of warranties with every copy of the Work You Distribute or Publicly Perform. When You Distribute or Publicly Perform the Work, You may not impose any effective technological measures on the Work that restrict the ability of a recipient of the Work from You to exercise the rights granted to that recipient under the terms of the License. This Section 4(a) applies to the Work as incorporated in a Collection, but this does not require the Collection apart from the Work itself to be made subject to the terms of this License. If You create a Collection, upon notice from any Licensor You must, to the extent practicable, remove from the Collection any credit as required by Section 4(c), as requested. If You create an Adaptation, upon notice from any Licensor You must, to the extent practicable, remove from the Adaptation any credit as required by Section 4(c), as requested.

b. You may Distribute or Publicly Perform an Adaptation only under the terms of: (i) this License; (ii) a later version of this License with the same License Elements as this License; (iii) a Creative Commons jurisdiction license (either this or a later license version) that contains the same License Elements as this License (e.g., Attribution-ShareAlike 3.0 US); (iv) a Creative Commons Compatible License. If you license the Adaptation under one of the licenses mentioned in (iv), you must comply with the terms of that license. If you license the Adaptation under the terms of any of the licenses mentioned in (i), (ii) or (iii) (the “Applicable License”), you must comply with the terms of the Applicable License generally and the following provisions: (I) You must include a copy of, or the URI for, the Applicable License with every copy of each Adaptation You Distribute or Publicly Perform; (II) You may not offer or impose any terms on the Adaptation that restrict the terms of the Applicable License or the ability of the recipient of the Adaptation to exercise the rights granted to that recipient under the terms of the Applicable License; (III) You must keep intact all notices that refer to the Applicable License and to the disclaimer of warranties with every copy of the Work as included in the Adaptation You Distribute or Publicly Perform; (IV) when You Distribute or Publicly Perform the Adaptation, You may not impose any effective technological measures on the Adaptation that restrict the ability of a recipient of the Adaptation from You to exercise the rights granted to that recipient under the terms of the Applicable License. This Section 4(b) applies to the Adaptation as incorporated in a Collection, but this does not require the Collection apart from the Adaptation itself to be made subject to the terms of the Applicable License.

c. If You Distribute, or Publicly Perform the Work or any Adaptations or Collections, You must, unless a request has been made pursuant to Section 4(a), keep intact all copyright notices for the Work and provide, reasonable to the medium or means You are utilizing: (i) the name of the Original Author (or pseudonym, if applicable) if supplied, and/or if the Original Author and/or Licensor designate another party or parties (e.g., a sponsor institute, publishing entity, journal) for attribution (“Attribution Parties”) in Licensor’s copyright notice, terms of service or by other reasonable means, the name of such party or parties; (ii) the title of the Work if sup-

plied; (iii) to the extent reasonably practicable, the URI, if any, that Licensor specifies to be associated with the Work, unless such URI does not refer to the copyright notice or licensing information for the Work; and (iv) , consistent with Section 3(b), in the case of an Adaptation, a credit identifying the use of the Work in the Adaptation (e.g., “French translation of the Work by Original Author,” or “Screenplay based on original Work by Original Author”). The credit required by this Section 4(c) may be implemented in any reasonable manner; provided, however, that in the case of a Adaptation or Collection, at a minimum such credit will appear, if a credit for all contributing authors of the Adaptation or Collection appears, then as part of these credits and in a manner at least as prominent as the credits for the other contributing authors. For the avoidance of doubt, You may only use the credit required by this Section for the purpose of attribution in the manner set out above and, by exercising Your rights under this License, You may not implicitly or explicitly assert or imply any connection with, sponsorship or endorsement by the Original Author, Licensor and/or Attribution Parties, as appropriate, of You or Your use of the Work, without the separate, express prior written permission of the Original Author, Licensor and/or Attribution Parties.

d. Except as otherwise agreed in writing by the Licensor or as may be otherwise permitted by applicable law, if You Reproduce, Distribute or Publicly Perform the Work either by itself or as part of any Adaptations or Collections, You must not distort, mutilate, modify or take other derogatory action in relation to the Work which would be prejudicial to the Original Author's honor or reputation. Licensor agrees that in those jurisdictions (e.g. Japan), in which any exercise of the right granted in Section 3(b) of this License (the right to make Adaptations) would be deemed to be a distortion, mutilation, modification or other derogatory action prejudicial to the Original Author's honor and reputation, the Licensor will waive or not assert, as appropriate, this Section, to the fullest extent permitted by the applicable national law, to enable You to reasonably exercise Your right under Section 3(b) of this License (right to make Adaptations) but not otherwise.

#### **5. Representations, Warranties and Disclaimer**

UNLESS OTHERWISE MUTUALLY AGREED TO BY THE PARTIES IN WRITING, LICENSOR OFFERS THE WORK AS-IS AND MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND CONCERNING THE WORK, EXPRESS, IMPLIED, STATUTORY OR OTHERWISE, INCLUDING, WITHOUT LIMITATION, WARRANTIES OF TITLE, MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, NON-INFRINGEMENT, OR THE ABSENCE OF LATENT OR OTHER DEFECTS, ACCURACY, OR THE PRESENCE OF ABSENCE OF ERRORS, WHETHER OR NOT DISCOVERABLE. SOME JURISDICTIONS DO NOT ALLOW THE EXCLUSION OF IMPLIED WARRANTIES, SO SUCH EXCLUSION MAY NOT APPLY TO YOU.

**6. Limitation on Liability.** EXCEPT TO THE EXTENT REQUIRED BY APPLICABLE LAW, IN NO EVENT WILL LICENSOR BE LIABLE TO YOU ON ANY LEGAL THEORY FOR ANY SPECIAL, INCIDENTAL, CONSEQUENTIAL, PUNITIVE OR EXEMPLARY DAMAGES ARISING OUT OF THIS LICENSE OR THE USE OF THE WORK, EVEN IF LICENSOR HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

#### **7. Termination**

a. This License and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this License. Individuals or entities who have received Adaptations or Collections from You under this License, however, will not have their licenses terminated provided such individuals or entities remain in full compliance with those licenses. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this License.

b. Subject to the above terms and conditions, the license granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different license terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this License (or any other license that has been, or is required to be, granted under the terms of this License), and this License will continue in full force and effect unless terminated as stated above.

#### **8. Miscellaneous**

a. Each time You Distribute or Publicly Perform the Work or a Collection, the Licensor offers to the recipient a license to the Work on the same terms and conditions as the license granted to You under this License.

- b. Each time You Distribute or Publicly Perform an Adaptation, Licensors offers to the recipient a license to the original Work on the same terms and conditions as the license granted to You under this License.
- c. If any provision of this License is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this License, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.
- d. No term or provision of this License shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.
- e. This License constitutes the entire agreement between the parties with respect to the Work licensed here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensors shall not be bound by any additional provisions that may appear in any communication from You. This License may not be modified without the mutual written agreement of the Licensors and You.
- f. The rights granted under, and the subject matter referenced, in this License were drafted utilizing the terminology of the Berne Convention for the Protection of Literary and Artistic Works (as amended on September 28, 1979), the Rome Convention of 1961, the WIPO Copyright Treaty of 1996, the WIPO Performances and Phonograms Treaty of 1996 and the Universal Copyright Convention (as revised on July 24, 1971). These rights and subject matter take effect in the relevant jurisdiction in which the License terms are sought to be enforced according to the corresponding provisions of the implementation of those treaty provisions in the applicable national law. If the standard suite of rights granted under applicable copyright law includes additional rights not granted under this License, such additional rights are deemed to be included in the License; this License is not intended to restrict the license of any rights under applicable law.

Creative Commons is not a party to this License, and makes no warranty whatsoever in connection with the Work. Creative Commons will not be liable to You or any party on any legal theory for any damages whatsoever, including without limitation any general, special, incidental or consequential damages arising in connection to this license. Notwithstanding the foregoing two (2) sentences, if Creative Commons has expressly identified itself as the Licensors hereunder, it shall have all rights and obligations of Licensors.

Except for the limited purpose of indicating to the public that the Work is licensed under the CCPL, Creative Commons does not authorize the use by either party of the trademark "Creative Commons" or any related trademark or logo of Creative Commons without the prior written consent of Creative Commons. Any permitted use will be in compliance with Creative Commons' then-current trademark usage guidelines, as may be published on its website or otherwise made available upon request from time to time. For the avoidance of doubt, this trademark restriction does not form part of the License.

Creative Commons may be contacted at <http://creativecommons.org/>.

# **Natural Language Processing, Multilinguality**

Editors Daniela Majchráková, Radovan Garabík

Cover Design by Vladimír Benko  
Typeset by Marek Kukan

Printed and published by Tribun EU s. r. o.  
Cejl 32, 602 00 Brno, Czech Republic

180 pages

First edition at Tribun EU  
Brno 2011

ISBN 978-80-263-0049-6

*www.librix.eu*

