В. Бенко
V. Benko

ЭКСПЕРИМЕНТ СО СОВМЕСТИМОЙ СКЕТЧ-ГРАММАТИКОЙ

COMPATIBLE SKETCH GRAMMAR EXPERIMENT

Аннотация. Наша статья описывает эксперимент переноса скетч-грамматики, первоначально созданной для словацких корпусов, на русский корпус, чтобы определить степень параллелизма между двумя скетчами. В рамках эксперимента мы создали и аннотировали русский веб-корпус среднего размера и обработали его с помощью системы Word Sketch Engine.

Abstract. Our paper describes an experiment of porting a sketch grammar originally created for Slovak corpora to a Russian corpus to determine the extent of parallelism in the resulting word sketches. In the framework of the experiment, we created and PoS-tagged a medium-size Russian corpus that was subsequently processed by the Word Sketch Engine.

1. Introduction

Word sketches are one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour[1]. To create word sketches we need a PoS-tagged corpus, the Word Sketch Engine (WSE) software and a set of rules – the sketch grammar.

In our Institute, the WSE has been extensively used since autumn 2007 with several Slovak and Czech corpora. These corpora serve as a source of lexical evidence for our monolingual and bilingual lexicographic projects[2], as well as for other linguistic research activities. The following text will show the results of an experiment to include a new language – Russian – into our set of corpora.

---

[1] *Kilgarriff A., Rychly P., Smrz P., Tugwell D.* The Sketch Engine. Proc. EURALEX 2004, Lorient, France; 2004. P. 105–116.

[2] *Benko V.* Optimizing Word Sketches for a large-scale lexicographic project. **http://videolectures.net/korpusi2010_benko_ows**

2. The Corpus

To create a test web corpus, we decided to use the recently released open-source tool SpiderLing[3] created at Masaryk University in Brno. SpiderLing has been optimized for collecting textual data from the web and contains an integrated language recognition module and a tool for boilerplate removal (jusText)[4]. The 100 input seed URLs were collected by means of Google. SpiderLing was started with three parallel processes (on a quad-core server) and, during its two runs with a total crawling time of just 14 hours, a total of 7.9 GB of raw text data was obtained.

3. PoS Tagging

The PoS tagging was performed by means of Helmut Schmid's TreeTagger[5] with the Russian parameter file trained by Serge Sharoff using the MULTEXT-East tagset[6]. This step took 39 hours to complete. The resulting corpus was subsequently filtered to fix some errors in tokenization of punctuation, to segment the sentences, and to de-duplicate it on the document level. After this processing, the corpus contained 799 million tokens in 604 thousand documents. As the last processing step, the corpus was de-duplicated at the sentence level using the fingerprint method. The duplicate sentences were not deleted from the corpus but rather they were marked by a special flag to exclude them from the WSE processing. In total, there were 27.6% duplicate sentences containing 22.1% of the corpus tokens.

Before starting work on the sketch grammar, we decided to unify the tags for punctuation as TreeTagger assigned three different tags for this purpose.

---

[3] *Suchomel V., Pomikálek J.* Efficient Web Crawling for Large Text Corpora. 7th Web as Corpus Workshop (WAC-7), Lyon, France; April 2012

[4] *Pomikálek J.* 2011. Removing Boilerplate and Duplicate Content from Web Corpora. Ph.D. thesis, Masaryk University, 2011.

[5] **http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/treetagger.html**

[6] **http://corpus.leeds.ac.uk/mocky**

4. The Sketch Grammar

The sketch grammar used in our WSE installation differs from most "traditional" grammars for corpora stored at sketchengine.co.uk[7] in several aspects:

- The rule names are not motivated syntactically (i.e., they do not indicate the syntactic relationship between the keyword and the collocate) but rather collocationally
- The right-hand or left-hand position of the collocate towards the keyword is indicated explicitly in the rule name
- The keyword's PoS in the rule is not specified, i.e., it covers any PoS
- Recall is preferred over precision

The object names within the rules are governed by the following rules:

- The keyword is denoted by the *X* symbol
- The keyword's grammatical attributes (mostly in unary rules) are indicated by lowercase abbreviation, e.g., *gen(X)* indicates the genitive case of keyword
- The collocate's PoS is indicated by an abbreviation with a leading capital letter, e.g., *Aj X* indicates a left-hand adjective collocate
- *Y* indicates a collocate that is from any PoS category
- *Z* indicates a collocate from any PoS category not covered by the other «explicit» rules

This way of writing sketch grammar has (against the traditional one) several advantages that can be conveniently utilized within the lexicographic projects:

- The rules symmetrically cover collocational relationships among all word classes, i.e., not only those where a direct syntactic relationship can be seen
- The rule names are more user-friendly
- The rules also cover situations where an incorrect morphological tag has been assigned to the keyword

[7] **http://trac.sketchengine.co.uk/wiki/SkE/CorpusQuerying**

5. Rule Name Summary

The core of our grammar consists of rules covering four basic autosemantic word classes. Taking into account our experience with early versions of the grammar, the rules for verbs (*Vb X/X Vb*) and adverbs (*Av X/X Av*) do not distinguish the left and right position of the respective collocate.

For nouns («substantives»), two separate rules take into account the position of the collocate (*Sb X, X Sb*). Similar situations can be found with adjectives (*Aj X, X Aj*) and prepositions (*Pp X, X Pp*).

The «catch all» rules for the remaining word classes (*Z X, X Z*) cover mostly numerals and pronouns, as well as some synsemantic word classes.

The remaining two binary (symmetric) rules map the relationship of coordination, i.e., the situation where a keyword and a collocate with an identical morphological tag are separated by a comma (*X/Y , X/Y*) or a conjunction (*X/Y Cj X/Y*).

The four trinary rules cover relationships among a keyword, collocate, and preposition in different positions (*Pp Y X*, *Pp X Y*, *Y Pp X*, and *X Pp Y*).

Our set of rules is complemented by unary rules showing the frequency distribution of the keyword's forms according to grammatical categories.

6. Compatible Sketch Grammars

If sketch grammars for corpora of different languages are written within the same paradigm, the resulting word sketches can be conveniently used in bilingual lexicography as a supplement to (or a partial replacement of non-existent) parallel corpora. This method is used in creating the new Czech-Slovak dictionary currently being compiled at our Institute. As Czech and Slovak are both linguistically and culturally closely related languages, the frequencies of collocates for a keyword are mostly very similar. Fig. 1 shows an example of such a “parallel” sketch derived from two comparable web corpora for the keyword *jazyk* «language». The second columns in both tables

show the raw frequency and the third ones the statistical value of salience[8].

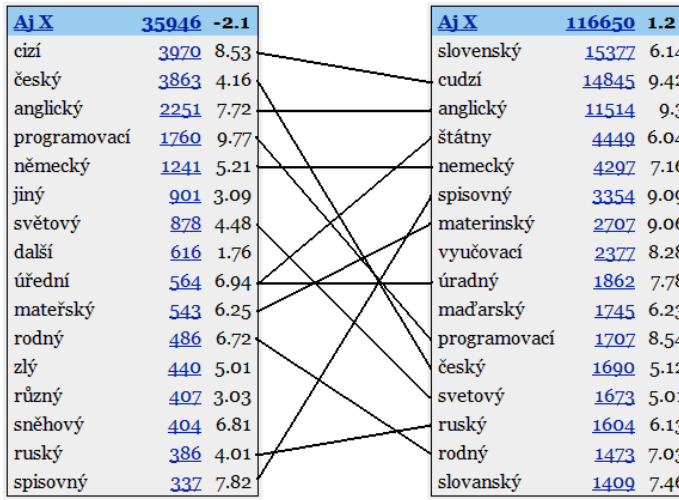


Fig. 1. Jazyk (cs) / jazyk (sk)

It can be seen that out of the 16 most frequent adjectival collocates on the source language (SL) side, 11 have their corresponding translation equivalents among the most frequent ones on the target language (TL) side (indicated by a line).

In compiling a bilingual dictionary, such parallel sketches can be effectively used for selection of appropriate examples. The author usually considers both the collocations appearing in both tables and also the most significant collocations that differ between the SL and the TL.

7. The Russian-Slovak Parallel Sketches

As the MULTEXT-East tagset used in PoS-tagging of the Russian corpus is quite similar to the Slovak tagset[9], porting the

[8] **https://trac.sketchengine.co.uk/attachment/wiki/SkE/DocsIndex/ske-stat.pdf?format=raw**

sketch grammar was a fairly straightforward venture, partially repeating the work done by Maria Khokhlova back in 2010[10].

Now, let us have a look at an example of the degree of parallelism that can be found between the Russian and Slovak word sketches. Within the limited size of this paper, we can only afford to present a small probe, covering just the already mentioned keyword *язык* «language». The Fig. 2 shows the situation of adjectival collocates.

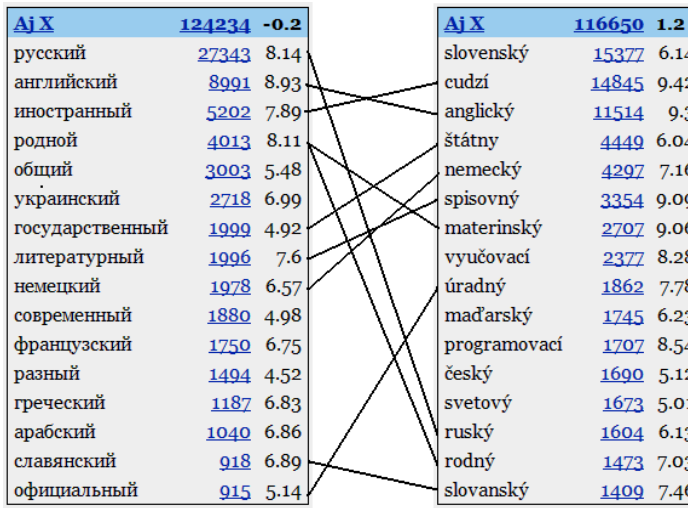


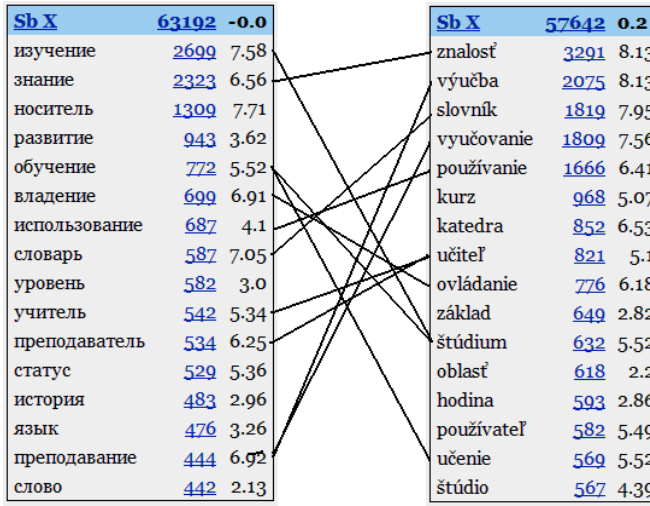
Fig. 2. Язык (ru) / jazyk (sk)

Though the number of parallelisms is smaller than that of Czech vs. Slovak, most SL collocates (9 out of 16) have TL translation equivalents. A similar situation can be observed with nouns in Fig. 3 (9 out of 16) and even better with verbs in Fig. 4 (12 out of 16). The final figure, Fig.5, shows some parallelisms seen within coordination collocates (8 out of 16).

---

[9] **http://korpus.juls.savba.sk/morpho.html**

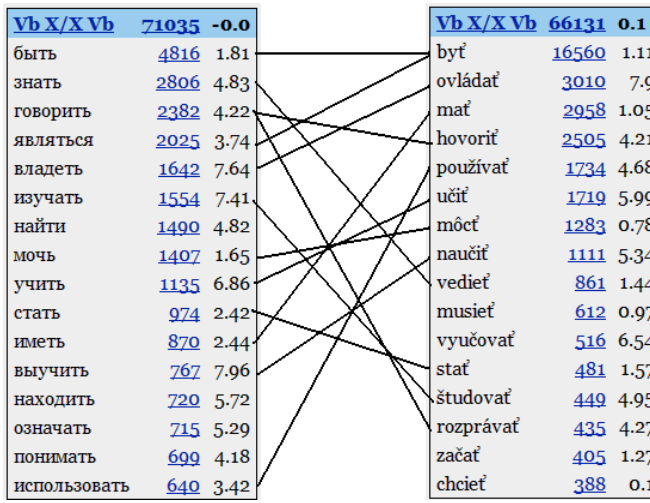
[10] *Khokhlova M.* Building Russian Word Sketches as Models of Phrases. Proc. EURALEX, Leeuwarden, July 2010.



Sb X	63192	-0.0
изучение	2699	7.58
знание	2323	6.56
носитель	1309	7.71
развитие	943	3.62
обучение	772	5.52
владение	699	6.91
использование	687	4.1
словарь	587	7.05
уровень	582	3.0
учитель	542	5.34
преподаватель	534	6.25
статус	529	5.36
история	483	2.96
язык	476	3.26
преподавание	444	6.92
слово	442	2.13

Sb X	57642	0.2
znalosť	3291	8.13
výučba	2075	8.13
slovník	1819	7.95
vyučovanie	1809	7.56
používanie	1666	6.41
kurz	968	5.07
katedra	852	6.53
učiteľ	821	5.1
ovládanie	776	6.18
základ	649	2.82
štúdium	632	5.52
oblasť	618	2.2
hodina	593	2.86
používateľ	582	5.49
učenie	569	5.52
štúdio	567	4.39

Fig. 3. Язык (ru) / jazyk (sk)



Vb X/X Vb	71035	-0.0
быть	4816	1.81
знать	2806	4.83
говорить	2382	4.22
являться	2025	3.74
владеть	1642	7.64
изучать	1554	7.41
найти	1490	4.82
мочь	1407	1.65
учить	1135	6.86
стать	974	2.42
иметь	870	2.44
выучить	767	7.96
находить	720	5.72
означать	715	5.29
понимать	699	4.18
использовать	640	3.42

Vb X/X Vb	66131	0.1
byť	16560	1.11
ovládať	3010	7.9
mať	2958	1.05
hovoriť	2505	4.21
používať	1734	4.68
učiť	1719	5.99
môcť	1283	0.78
naučiť	1111	5.34
vedieť	861	1.44
musieť	612	0.97
vyučovať	516	6.54
stať	481	1.57
študovať	449	4.95
rozprávať	435	4.27
začať	405	1.27
chcieť	388	0.1

Fig. 4. Язык (ru) / jazyk (sk)

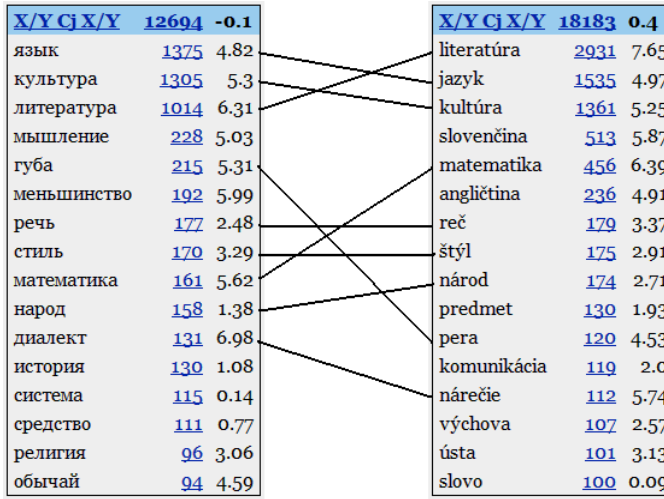


Fig. 5. Язык (ru) / jazyk (sk)

8. Conclusion and Further Work

Our experiment was targeted both at testing the new tool for creating web corpora, SpiderLing, and at finding out whether the degree of parallelism between Czech and Slovak word sketches can also be observed between Russian and Slovak, if a compatible sketch grammar is used. As all the tools necessary were either freely available or could be easily modified from our own tools, the whole experiment took approximately only 2 weeks to complete. SpiderLing proved to be extremely effective in gathering Russian web data. Initial analyses of the sketch data show that the degree of parallelism between Russian and Slovak is similar to that between Czech and Slovak and their use for a bilingual lexicographic project is feasible.Our experiment will continue in two directions. We want to increase the size of the Russian corpus to make it comparable with the largest Slovak corpus available. We also want to include other European languages into our corpus set, at least those covered by TreeTagger.