Published in: Bański, Piotr/Kupietz, Marc/Lüngen, Harald/Rayson, Paul/Biber, Hanno/Breiteneder, Evelyn/Clematide, Simon/Mariani, John/stevenson, Mark/Sick, Theresa (eds.): Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017 including the papers from the Web-as-Corpus (WAC-XI) guest section. Birmingham, 24 July 2017. - Mannheim: Institut für Deutsche Sprache, 2017. Pp. 43-48

Are Web Corpora Inferior? The Case of Czech and Slovak

Vladimír Benko
Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics
Panská 26, SK-81101 Bratislava
and
Comenius University in Bratislava
UNESCO Chair in Plurilingual and Multicultural Communication
Šafárikovo nám. 6, SK-81499 Bratislava
vladob@juls.savba.sk

Abstract

Our paper describes an experiment aimed to assessment of lexical coverage in web corpora in comparison with the traditional ones for two closely related Slavic languages from the lexicographers' perspective. The preliminary results show that web corpora should not be considered "inferior", but rather "different".

1 Introduction

During the last 15 years, creation of web corpora has been recognized as an effective way of obtaining language data in situations where building traditional corpora would be either too costly or too slow (Baroni et al., 2009; Jakubíček et al., 2013; Schäfer & Bildhauer, 2013) and building and analyzing web corpora has transformed into a separate branch of corpus linguistics.

At present, both traditional and web corpora do exist for many languages, with the respective web corpus being of comparable or even larger size. Any (corpus) linguist in this situation is therefore confronted with questions as follows: How does the existence of two "language samples" created by different methodology and technology influence my linguistic research? Which corpus provides better evidence allowing for generalizing my conclusions? Is any of the corpora "infereior"?

Both Czech and Slovak belong to languages where we can try looking for answers to such questions as respective corpora exist and the source data is (in our case) available.

2 Comparing Corpora

Due to the huge sizes of contemporary corpora, any comparison of their contents is a challenging task. For corpora available on-line, some comparisons can be performed via the respective interface, optionally in combination with the frequency lists generated from the respective corpora (Khokhlova, 2016). The large-scale statistical evaluation, however, requires having the source corpus data available (Kilgarriff, 2001).

Besides the assessment of lexical coverage based on rank and frequency distributions of word forms and/or lemmas, other corpus properties may also be compared, e.g. the "quality" of morphosyntactic annotation (out-of-vocabulary rate), "noise" (undetected foreign language and/or duplicate text fragments). If a tool for collocational analysis is available, such as Sketch Engine (Kigarriff et al., 2004; Kilgarriff et al., 2014), collocation profiles for a selected set of keywords can be conveniently compared.

3 The Experiment

In our paper, we describe an on-going experiment, in the framework of which we try to evaluate the lexical coverage of web corpora in comparison with the traditional corpora for the respective languages. As our comparison is mainly motivated by the needs of lexicographers, in an ideal case, it would be useful to compare the proportion of lexical items found in the respective corpora and not covered by existing dictionaries, that would qualify to become headwords in a newly complied dictionary (e.g., neologisms).

Such a task, however, would involve a lot of manual work – it is not enough just to count "out-of-vocabulary" tokens derived from the respective corpora: the web corpus naturally contains more of them because of more "noise".

We have therefore decided to do something that can be performed without any manual evaluation. The procedure involved comparing frequency lists derived from the respective corpora with headword lists of medium-sized dictionaries. As we were also interested how the corpus size influences the lexical coverage, we performed the same experiment with subcorpora of various sizes created by (random) sampling of the respective traditional and web corpus data.

3.1 The corpora

The traditional Czech corpora were represented by the *syn* series of the Czech National corpus (Křen et al., 2014) available from the LINDAT portal. The "opportunistic" *syn v4* basically contains all Czech corpus data gathered by the Institute of Czech National Corpus, making it rather unbalanced. A well-balanced part (containing the four representative 100 Megaword Czech corpora, i.e., *syn2000*, *syn2005*, *syn2010* and *syn2015*, respectively), however, can be easily extracted from *syn v4* by means of its metadata, yielding a balanced 400+ Megaword corpus that will be referred to as *syn20xx*.

The Slovak traditional corpora were represented by the *prim* series of the Slovak National Corpus (Šimková – Garabík, 2014; SNK, 2015). Two subcorpora have been used in our research – the 835 Megaword unbalanced *prim-6.1-all* (SNK, 2013a), and the 300+ Megaword balanced *prim-6.1-vyv* (SNK, 2013b). The source data of these corpora are, unfortunately, not available for users outside of our Institute.

The web corpora have been represented by the *Maximum* class of the Aranea Project corpora (Benko, 2014), i.e., the 5+ Gigaword *Araneum Bohemicum* for Czech, and the 3+ Gigaword *Araneum Slovacum* for Slovak.

To ensure the maximal compatibility of annotation among the corpora, both Czech and Slovak traditional corpora have been retokenized and retagged before being used in our experiment, which resulted in slight decrease of their original size measured in tokens. The information on corpora is summarized in Table 1.

Name	Language	Type	Size
syn 20xx	Czech	traditional, balanced	462 M tokens
syn v4	Czech	traditional	4,352 M tokens
Araneum Bohemicum Maximum (BM)	Czech	web	5,174 M tokens
prim-6.1-public-vyv	Slovak	traditional, balanced	317 M tokens
prim-6.1-public-all	Slovak	traditional	858 M tokens
Araneum Slovacum III Maximum (SM)	Slovak	web	3,357 M tokens

Table 1. Corpora used

3.2 Sampling Subcorpora

The subcorpora used in our experiment have been sampled in a logarithmic scale graded as follows: *1M*, *2M*, *5M*, *10M*, *20M*, …, etc., up to the actual corpus size. The rudimentary sampling algorithm was based on splitting each 1-Megaword block into two parts defined by the parameter. Though this procedure can be considered "radom" for very large subcorpora, it is certainly not the case with the small ones.

For each subcorpus, a frequency list has been extracted containing both lemmas and word forms, accompanied by the PoS information.

3.3 The wordlists

The only relatively new Czech dictionary available in electronic form that could be used to extract the Czech wordlist for our experiment was the (retro-digitized) bilingual Czech-Slovak Dictionary (Horák et al. 1981). The situation has been more favorable for Slovak, where several dictionaries in electronic form were available. We have opted here for the dictionary part the Rules of the Slovak Orthography (PSP, 2000), as its size is on par with the Czech dictionary used.

The extracted headword lists have been filtered to get rid of multi-word expressions (mostly secondary prepositions and loanwords), and to remove reflexive formants "se/si" for Czech and "sa/si" for Slovak that appear as parts of headwords with reflexive verbs, but would not have a counterpart in wordlists derived from corpora. After this processing the Czech list contained approx. 73,500, and the Slovak list 65,500 headwords, respectively.

.

Corpus size (M)	syn 20xx			syn v4			Araneum BM		
	(1+)	(10+)	(100+)	(1+)	(10+)	(100+)	(1+)	(10+)	(100+)
1	32.13	8.77	1.42	32.97	8.84	1.35	31.92	8.91	1.48
2	40.07	13.31	2.78	39.23	13.03	2.66	39.39	13.10	2.77
5	51.54	22.49	5.56	47.91	20.63	5.48	49.51	20.93	5.48
10	59.38	30.46	8.86	54.65	27.22	8.47	57.20	28.37	8.49
20	65.79	38.89	13.59	61.03	34.64	12.54	64.03	36.43	12.84
50	74.93	51.36	22.95	68.48	44.38	19.91	71.85	46.99	20.67
100	79.65	59.08	31.44	73.73	51.89	26.80	76.57	54.61	27.75
200	83.05	66.09	40.27	77.80	58.42	34.14	80.52	61.64	35.78
(<) 500	86.06	73.56	50.39	82.44	66.70	44.03	84.48	70.09	46.62
1000				85.07	72.26	51.56	86.55	75.54	54.50
2000				86.96	77.15	58.49	87.92	79.85	61.62
(<) 5000				88.32	81.48	65.54	89.14	84.28	70.33

Tab. 2. Lexical Coverage for Czech

3.4 Processing the Czech Data

The proportion of the dictionary headword list (in %) covered by the respective subcorpus has been observed. All results are displayed in Table 2.

For each sampled subcorpus, three values are presented, representing the subcorpus lexical coverage of the dictionary headword list if at least one, ten, and one hundred occurrences of lexical items in the corpus are required, respectively. For example, the 100 M subcorpus sampled from syn 20xx covers 79.65% of the dictionary headword list on condition that 1 corpus occurrence is considered satisfactory, but only 31.44% if at least 100 corpus occurrences are required.

The values from the table are visualized in Fig. 1. The *x* axis represents the corpus size in millions of tokens and the *y* axis shows the coverage of vocabulary (in %) by the respective (sub)corpora. As the left part of the graph is rather dense, the situation with smaller subcorpora is better visible if corpus size is plotted in a logarithmic scale (Fig. 2).

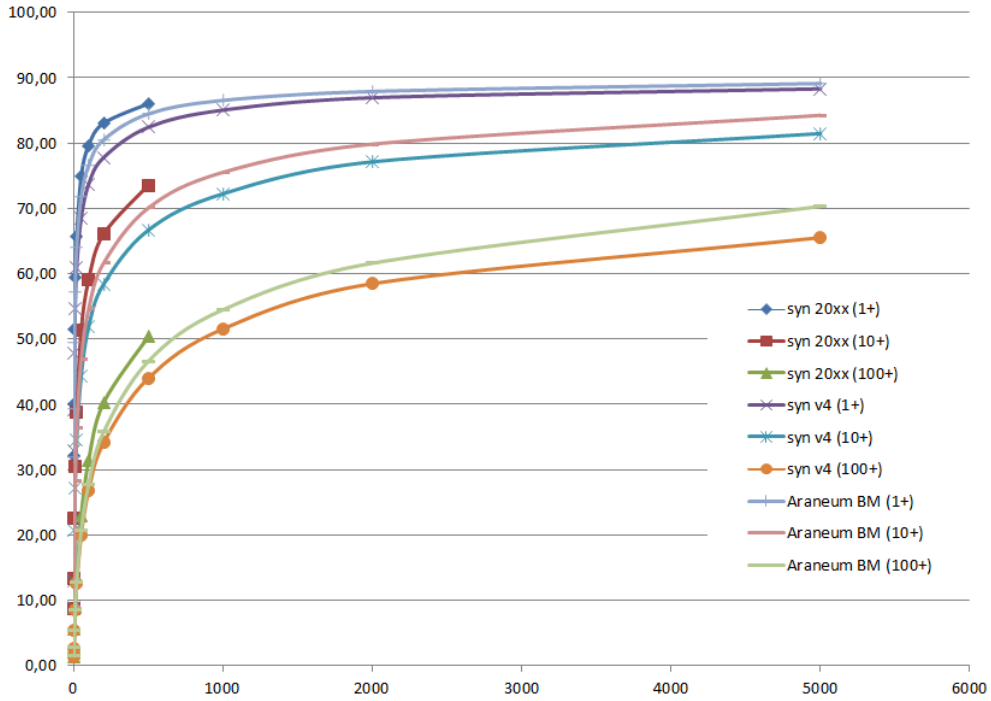

Fig. 1

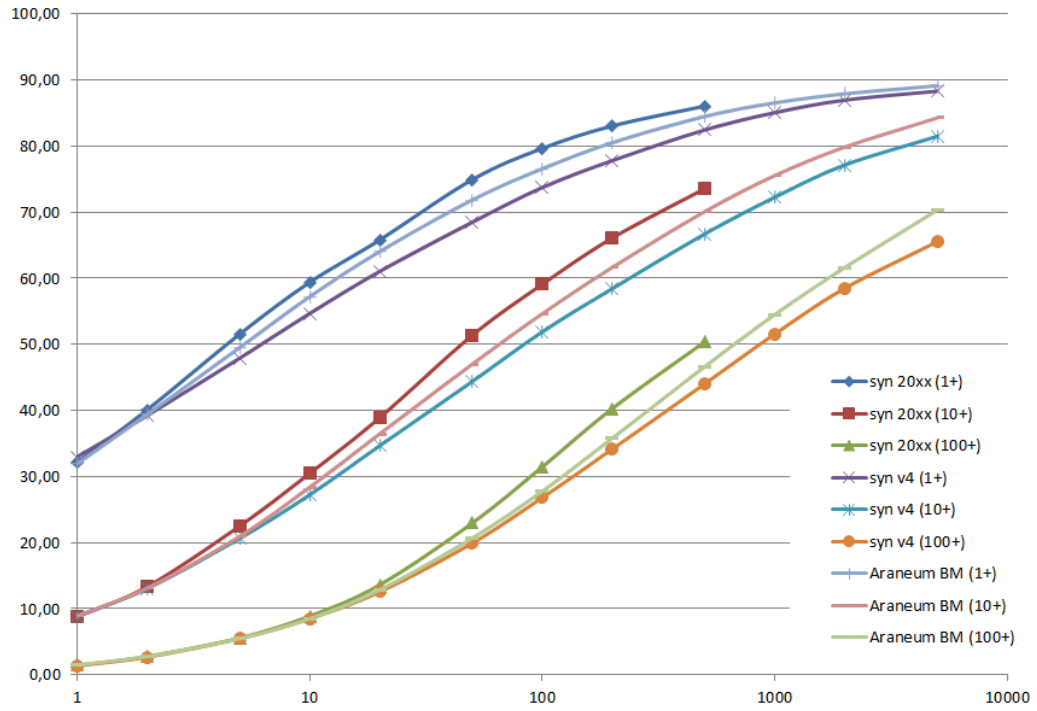

Fig. 2

3.5 The Slovak Data

The procedure for Slovak was similar to that of Czech, with the main difference being the sizes of both traditional and web corpora. The respective results are summarized in Table 3.

Corpus size (M)	pri,m-6.1-vyv			prim-6.1-all			Araneum SM		
	(1+)	(10+)	(100+)	(1+)	(10+)	(100+)	(1+)	(10+)	(100+)
1	31.66	8.38	1.26	31.47	8.59	1.30	30.50	8.69	1.41
2	39.30	13.06	2.50	38.08	12.57	2.54	37.84	12.93	2.72
5	52.00	22.43	5.26	49.16	20.58	5.14	48.62	20.60	5.29
10	59.66	30.32	8.62	56.58	28.03	8.18	55.74	27.59	8.33
20	66.69	39.13	13.51	64.67	36.64	12.54	62.10	35.61	12.57
50	74.64	51.63	22.85	73.05	49.08	21.22	69.55	45.77	20.11
100	78.62	59.97	31.25	77.28	56.84	29.00	74.31	53.19	27.09
200	81.58	66.88	40.23	80.76	64.30	37.45	77.99	59.90	35.00
(<) 500				83.68	72.07	48.94	81.76	68.04	45.37
(<) 1000				84.83	75.82	55.36	83.64	72.89	52.94
2000							84.98	76.02	58.12
3500							85.22	77.84	59.76

Tab. 3. Lexical Coverage for Slovak

The figures show similar progress as those for Czech, forming the shapes displayed at Fig. 3 (in linear scale for subcorpora sizes), and Fig. 4 (logarithmic scale).

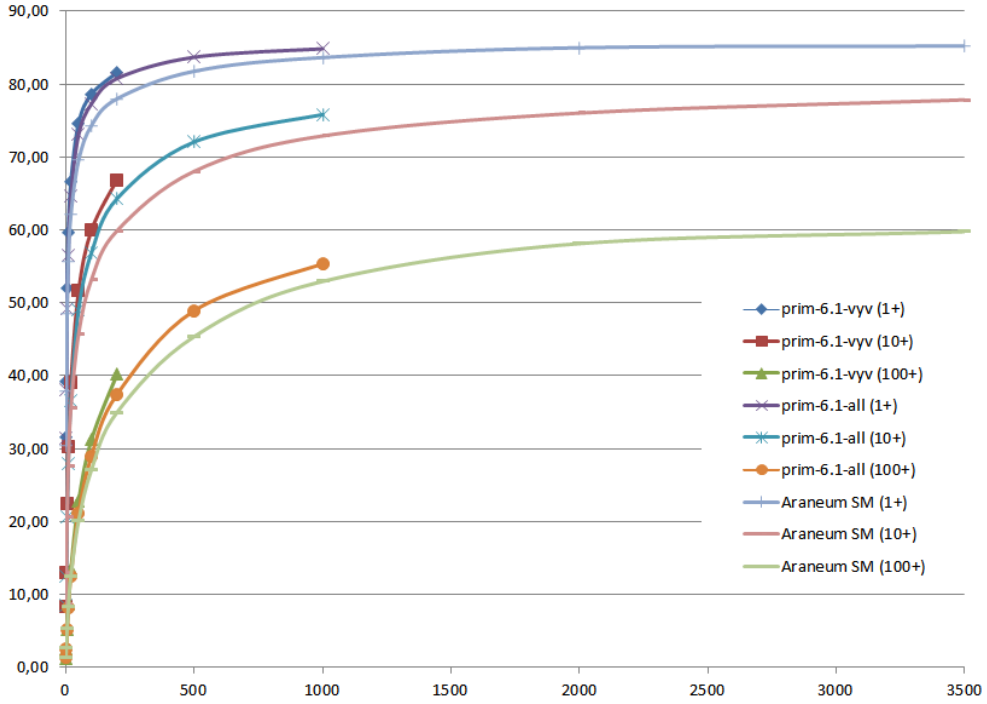

Fig. 3

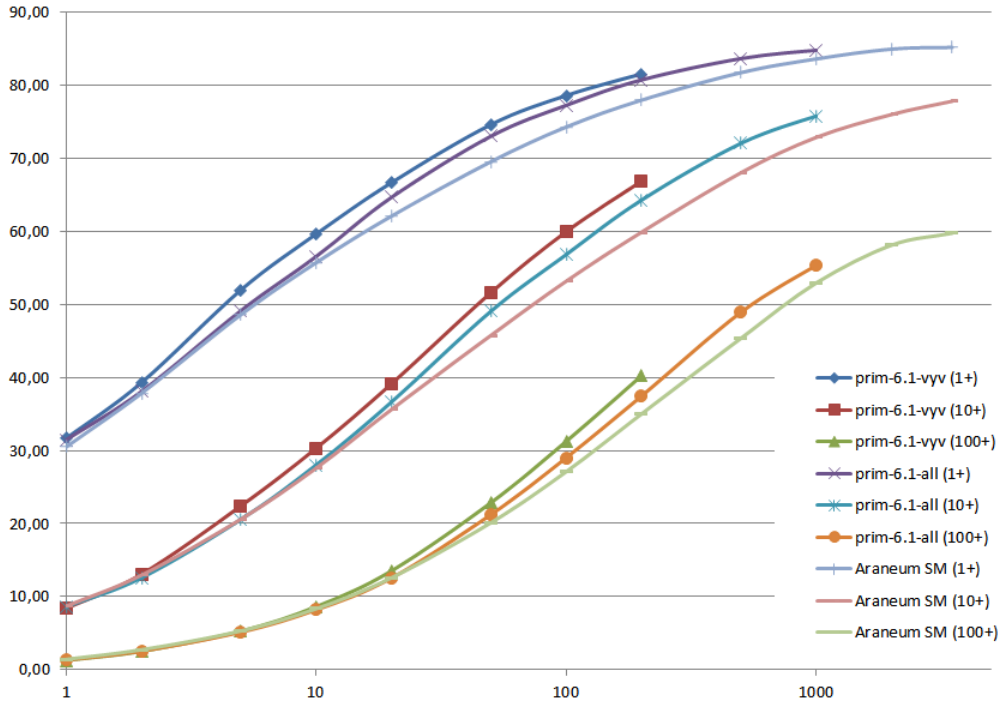

Fig. 4

4 Conclusion and Further Work

The results are mostly consistent with our expectations, and can be summarized as follows:

(1) The lexical coverage for both languages is growing steeply with the size of corpus for smaller corpora, but a saturation can observed at approximately 1 billion tokens.

(2) The coverage of the Czech headword list approaches 90%, while the Slovak one stops at approximately 85%, which deserves a more detailed analysis. The quick lookup reveals several cases here: the Czech headword lists contained many regular derivates from infrequent words, spelling variants not present in contemporary language, and even typos in the retro-digitized dictionary); the unmatched items in the Slovak list also contain a large number in geographical and inhabitant names that rarely occur in text.

(3) Both balanced corpora are slightly "better" within the range of their size, this advantage can be outperformed be the sheer size of larger corpora.

(4) Traditional unbalanced corpus is slightly "worse" in smaller sizes for Czech and slightly "better" for Slovak. The difference, however, almost disappears with corpora larger than 2 billion tokens.

(5) As a source for lexicographic work, (at least) 2 Gigaword corpus is to be recommended.

More research is necessary to evaluate the differences between traditional and web corpora, most notably in text types, domains, genres and registers, as well as with wordlist derived from different dictionaries.

Acknowledgement

The research described in this paper has been supported by VEGA Grant Agency, Project No. 2/0017/17.

References

Baroni M., Bernardini, S., Ferraresi A., Zanchetta E. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. Language Resources and Evaluation 43 (3), pp. 209–226.

Benko, V. 2014. Aranea: Yet another Family of (Comparable) u Corpora. In Text, Speech, and Dialogue. 17th International Conference, TSD 2014 Brno, Czech Republic, September 8 – 12, 2014, Proceedings. Ed. P. Sojka et al. – Cham – Heidelberg – New York – Dordrecht – London : Springer, 2014, 21–29. ISBN 978-3-319-10816-2.

Jakubíček, M. – Kilgarriff, A. – Kovář, V. – Rychlý, P. – Suchomel V. 2013. The TenTen Corpus Family. In 7th International Corpus Linguistics Conference, Lancaster, July 2013.

Horák, G. et al. (Ed.) 1981, Česko-slovenský slovník. Bratislava : Veda, 1981.

Khokhlova, M. 2016. Large Corpora and Frequency Nouns. In Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016", Moscow, June 1–4, 2016.

Kilgarriff, A. 2001. Comparing Corpora. International Journal of Corpus Linguistics, 6(1), 97–133.

Kilgarriff, A. et al. 2004. The Sketch Engine. In: G. Williams and S.Vessier (eds.), Proceedings of the eleventh EURALEX International Congress EURALEX 2004 Lorient, France, July 6-10, 2004. Lorient : Université de Bretagne-Sud, pp. 105–116.

Kilgarriff, A., Rychlý, P., Jakubíček, M., Kovář, V., Baisa, V., Kocincová, L. 2014. Extrinsic Corpus Evaluation with a Collocation Dictionary Task. Proc. LREC 2014. Reykjavik, pp. 545-552.

Křen, M., Cvrček, V., Čapka, T. et al., 2016, SYN v4: large corpus of written Czech, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, http://hdl.handle.net/11234/1-1846.

PSP 2000. Považaj, M. (Ed.): Pravidlá slovenského pravopisu. 3., upravené a doplnené vydanie. Bratislava : Veda 2000.

Schäfer, R. – Bildhauer, F. 2013. Web Corpus Construction. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Šimková, M. – Garabík, R.: Slovenský národný korpus (2002–2012): východiská, ciele a výsledky pre výskum a prax. In: Jazykovedné štúdie XXXI. Rozvoj jazykových technológií a zdrojov na Slovensku a vo svete (10 rokov Slovenského národného korpusu). Ed. Katarína Gajdošová — Adriána Žáková. Bratislava: VEDA 2014, 35– 64.

SNK 2013a. Slovenský národný korpus – prim-6.1-public-all. Bratislava: Jazykovedný ústav Ľ. Štúra SAV, http://korpus.juls.savba.sk.

SNK 2013b. Slovenský národný korpus – prim-6.1-public-vyv. Bratislava: Jazykovedný ústav Ľ. Štúra SAV, http://korpus.juls.savba.sk.