Intrinsic Methods for Comparison of Corpora

Vít Baisa and Vít Suchomel

Masaryk University,
Botanická 68a
Brno, Czech Republic
xbaisa@fi.muni.cz
xsuchom2@fi.muni.cz

Abstract. Since there are only very few techniques for quantitative and systematic comparison of text corpora we proposed and implemented several novel methods. The procedures were applied to comparing two very large web based Czech text corpora: czTenTen12 and Hector with more than 4.47 and 2.65 billion words, respectively. All methods are fully automatic and some of them are even language independent. We released some of them so they can be used instantly for comparison of other corpora.

Key words: text corpus, corpora comparison

1 Introduction

Nowadays, thousands of new corpora are built each month. using automatic methods like WebBootCaT [1] and similar. In some systems, creating a new corpus is a matter of several mouse clicks. But despite the overwhelming amount of corpora available now there is no method for their comparison.

It clearly depends on the purpose and usage of the corpus: sometimes, just the size matters, sometimes texts in colloquial or internet language are required etc. In this paper we describe intrinsic methods for comparing corpora.

We were interested especially in comparing two recent very large web-based Czech text corpora – czTenTen12 [2] and Hector [3,4].

Methods presented in this paper are divided into several groups: a) general intrinsic properties, b) text cleaning and processing, c) wordlist-based methods and d) syntactic analysis.

Some initiatives dealing with comparing corpora were [5,6] and [7] but in general not much attention was paid to this topic.

2 General intrinsic techniques

2.1 Size

The basic intrinsic measure is the size of the corpus (number of words or tokens in the data). Generally, the larger the corpus, the better: 'Most phenomena

Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2013, pp. 51–58, 2013.

in natural languages are distributed in accordance with Zipf's law, so many words, phrases and other items occur rarely and we need very large corpora to provide evidence about them' [8]. There are 1.71 times more words and 1.39 times more sentences in czTenTen12 than in Hector according to Table 1. The measurement of words, tokens and sentences depends on the means of tokenization and sentence detection algorithms used for processing corpus data. These algorithms may differ in various corpora.

Table 1: Comparison of corpus size – gigabytes of textual data, billions of tokens, billions of words, millions of sentences

CORPUS	BYTES	TOKENS	WORDS	SENTENCES
Hector	17 GB	3.285 bn	2.607 bn	219 m
czTenTen12	31 GB	5.437 bn	4.458 bn	303 m

2.2 Diversity of sources

Unlike czTenTen12, Hector was constructed from manually selected web sites with large and good-enough-quality textual content (e.g. news servers, blog sites, discussion fora) [4]. Although such selection of particular documents may contribute to text quality, it may also decrease text diversity, e.g. genres like novels, legal documents or descriptions of goods are completely omitted. Therefore diversity of sources should be taken into account when building web corpora.

To measure the diversity of corpus sources, one could count number of web pages, web domains and top level domains represented in the corpus. The more diverse source of the data, the better coverage of language by the corpus may be expected. Since Hector was not available with necessary metadata, only the diversity of czTenTen12 could be evaluated and displayed in Table 2.

Spreading corpus sources over many top level domains may be useful for languages spoken all over the world (e.g. English) or to obtain variants of the language spoken in diferrent countries (e.g. Brasilian vs. European Portuguese). However, it may not help to find good quality texts in languages spoken in a single country like Czech.

Table 2: Diversity of sources – web pages, web domains, average number of pages per domain, median of pages per domain, top level domains

CORPUS	PAGES	DOMAINS	AVG	MED	TLDS
czTenTen12	9,747,315	233,122	42	4	97.6 % cz

3 Text processing and cleaning metrics

3.1 Sentence length

In the footsteps of [4] comparing sentence length of Czech corpora SYNT2005 and Hector, czTenTen12 is added to the comparison in Figure 1.

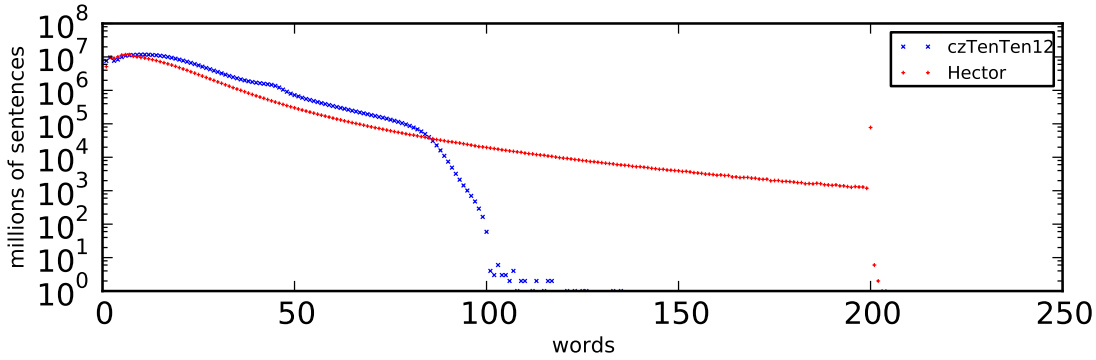

Fig. 1: Distribution of number of words in sentences (logarithmic y scale)

It can be seen the algorithm for detection of sentence borders trims sentences with more than 200 words in case of Hector. Another approach is used in case of czTenTen12 – the algorithm is more likely to end sentences longer than 80 words, see Table 3

Table 3: Sentence length metrics – peak length, average length, median length, observed threshold of breaking long sentences

CORPUS	PEAK	AVG	MED	LONG SENTENCES
Hector	7	15.0	12	hard limit of 200 words
czTenTen12	10	17.9	16	less strict rules after 80 words

3.2 Data duplicity

The less duplicate texts in a corpus the better. However, a very strict deduplication results in removing usable data needlessly. The deduplication strength of both examined corpora was intentionally selected by respective corpus designers. Duplicate and near duplicate texts were avoided in both corpora using a n-gram comparison method: paragraphs containing more than 30% seen 8-grams were removed from Hector, while paragraphs containing more than 50% seen 7-grams were removed from czTenTen12. Although both methods are similar, particular algorithms are different. We propose to compare the degree of deduplication based on a stricter n-gram comparison. Table 4 contains results of the

experiment performed using onion[1] set to remove sentences consisting of 50% seen 5-grams of sentences (smoothing disabled). Since the observed size drops are not large, it can be concluded both corpora are deduplicated sufficiently. CzTenTen12 was deduplicated more strictly than Hector.

Table 4: Corpus size difference after a strict deduplication of sentences

CORPUS	BYTES	TOKENS	SENTENCES
Hector	-23.3 %	-25.8 %	-23.6 %
czTenTen12	-17.6 %	-18.7 %	-18.4 %

4 Wordlist based techniques

4.1 The test

One of several steps in building a new corpus is language filtering. The aim is usually to have only one language in corpus so the language filter must identify texts out a desired language and remove them from the corpus. A problem might emerge if there are many small portions of foreign language texts below paragraph level. In this case one needs to set a level of granularity for the language filtering. But in general: the less words from a foreign language in your corpus the better filtered it is. Language statistics are not worsened by noise from a foreign language.

The test takes positions of all variants of English determiner *the* (THE, the, The, thE etc.) from a wordlist. These positions are then compared between examined corpora. The determiner was chosen since it is the most frequent word in English texts.

Table 5: The test results for Hector and czTenTen12

czTenTen12		**Hector**	
The	941	The	757
the	1,109	the	1,185
THE	24 k	THE	12 k
THe	942 k	ThE	264 k
tHe	2.4	tHe	314 k
ThE	2.7 M	THe	435 k
tHE	4.8 M	thE	654 k
thE	4.8 M	tHE	847 k

[1] http://nlp.fi.muni.cz/projects/onion

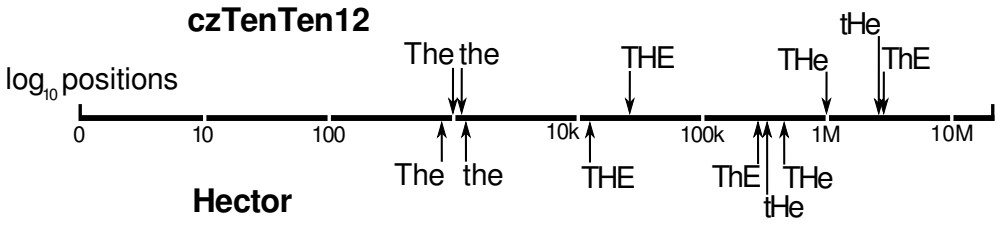

Fig. 2: Visualization of The test

You can see results in Table 5 and visualization in Figure 2. Most of variants of the determiner in Hector are more frequent then in czTenTen12 wordlist which could be interpreted as that czTenTen12 contains less portions of English texts.

4.2 Filtering wordlists

The motivation for this method is very similar to The test – we want to know if a corpus contains only the desired language. For this purpose we use morphologic analyzer to filter out all unknown words from wordlists and then check how many words remained.

It is not much important if the analyser can recognize all Czech words in wordlists. It is fair that the same filtering is applied on both corpora. We used Czech fast analyser Majka [9].

Results of this method also reveal problems in missing diacritics, wrong encoding of texts, number of typos – all these are not recognised by the analyzer and at the same time are not desirable in corpora.

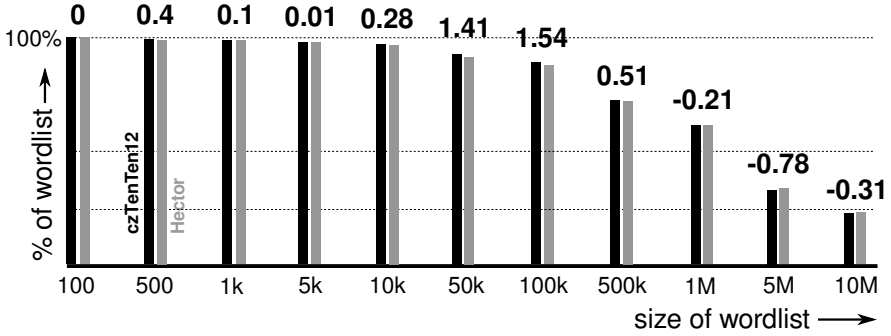

Fig. 3: Visualization of filtering wordlist test

In Figure 3 you can see results of filtering wordlists for Hector and czTenTen12. On x axis there are various lengths of wordlists: we filtered top parts of wordlists to see how the filtering is changing from top to bottom of wordlists.

The lengths of filtered resulting lists are very similar: the height of a rectangle is ratio between length of unfiltered and filtered list. The number

above each column is difference between ratio of czTenTen12 and Hector. If the number is positive then the respeective part of czTenTen12 filtered list is bigger then of the Hector's and means that Hector was filtered more.

First 500,000 words from both wordlists are less filtered in czTenTen12 but the rest of wordlists are less filtered in Hector. The interpretation of these results is not straightforward but we can conclude that from the perspective of this method both corpora are very similar and Hector is slightly better for low frequency words.

4.3 Keyword comparison

Following [7], we extracted lowercase keywords from czTenTen12 with Hector as the reference corpus (and vice versa) to explore in which words these corpora differ the most. It can be observed both recent web corpora contain more data from internet message boards and less news documents than the Czech National Corpus. In addition, there is much text from women fora in Hector.

Notable top keywords from Hector vs. czTenTen12:

- blog, holky, teda, taky, ahoj, fakt, ahojky, super, moc, ráda, takže – mostly informal, some in feminine gender (discussions of women)
- chtěla, řekla – verbs in feminine gender
- jdu, budu, máš, mám, jsi, nevím, doufám, jsem – 1st/2nd person (discussions)
- dneska, zítra, sem, teď, včera, pořád, tady, nějak – adverbs (blogs, discussions)

Notable top keywords from czTenTen12 vs. Hector:

- http, kdyz – poor tokenization, missing diacritics
- již, lze, dále, mohou, zejména, především – standard language (books, news)
- společnosti, oblasti, města, společnost, projektu, řízení, prostředí – society (news)
- zařízení, systém, nabízí, služby, informace – business
- této, tato, tyto, těchto, tohoto – demonstrative adj., standard language (news)

Notable top keywords vs. SYN2000 (Czech National Corpus) [10]:

- Hector vs. SYN2000: taky, teda, ahoj, holky, mám, fakt, moc, sem, dneska, takže, blog, nevím, máš, super, ráda, ahojky (discussions of women)
- czTenTen12 vs. SYN2000: taky, můžete, moc, děkuji, takže, cca, mám, dobrý, opravdu, dle, ahoj, bych, jestli, díky, hodně, super (discussions)
- SYN2000 vs. Hector and czTenTen12: praha, včera, korun, procent, české, vlády, státní, miliónů, zákona, trhu, ministr, ředitel, výstava, společnost, nato, prezident, čtk – standard language, news, Prague

5 Syntactic analysis

5.1 Syntactic functions

A good general corpus should consist mostly of syntactically correct sentences. It is not our intention to filter out syntactically problematic but otherwise quite common and understandable sentences. We aim to detect web garbage such as page navigation and labels, tables consisting of single words or numbers, computer program code samples, keywords used to increase page rank, link spam, artificially generated texts.

Let us declare a nice sentence contains the main syntactic roles – subject and predicate. Although this strict definition does not allow many correct sentences, it surely rules out unwanted content stated above.

Syntactic analysis tool Set [11] was used to carry out the experiment. Subsets of 15 million random sentences from examined corpora were syntactically tagged. Presence of subject and predicate was evaluated for each clause in all sentences. Table 6 reveals czTenTen12 contains slightly more sentences having the subject – predicate couple than Hector. A significant presence of web discussions in Hector is most likely the cause.

Table 6: Ratio of nice clauses in examined corpora – nice clauses (NCL), sentences with all clauses nice (NSEN), sentences with some but not all clauses nice (PNSEN)

CORPUS	NCL	NSEN	PNSEN
Hector	36.6 %	19.0 %	23.7 %
czTenTen12	39.6 %	23.6 %	29.2 %

6 Future work

We plan to implement other intrinsic methods using e.g. language models trained on different corpora: given a language model trained on corpus A we can then measure perplexity of the respective language model using corpus B and vice versa.

Another intrinsic methods to be developed are finding topics in corpora using available tools as e.g. Gensim [12] and measuring homogenity of corpora.

We have already tried some extrinsic methods for comparing corpora: one of them is Word sketch evaluation described in article *How to compare corpora* already submitted to LREC 2014 – the method is based on automatic extraction of good collocations from corpora.

We have also carried out extrinsic method described in [13] for these two corpora. Results will be published in a separate paper soon.

7 Conclusion

We described eight methods which can be used for a general systematic comparison of text corpora. We provided also results of these methods based on comparison of two very large Czech text corpora czTenTen12 and Hector.

The methods are ready to be used and you can download related tools and data from website of Natural Language Processing Centre.[2]

Acknowledgements This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013.

References

1. Baroni, M., Kilgarriff, A., Pomikálek, J., Rychlỳ, P.: Webbootcat: instant domain-specific corpora to support human translators. In: Proceedings of EAMT. (2006) 247–252
2. Suchomel, V.: Recent Czech web corpora. In Aleš Horák, P.R., ed.: 6th Workshop on Recent Advances in Slavonic Natural Language Processing, Brno, Tribun EU (2012) 77–83
3. Spoustová, J., Spousta, M., Pecina, P.: Building a web corpus of czech. (2010)
4. Spoustová, J., Spousta, M.: A high-quality web corpus of czech. In: LREC. (2012) 311–315
5. Kilgarriff, A.: Comparing corpora. International journal of corpus linguistics **6**(1) (2001) 97–133
6. Rayson, P., Garside, R.: Comparing corpora using frequency profiling. In: Proceedings of the workshop on Comparing Corpora, Association for Computational Linguistics (2000) 1–6
7. Kilgarriff, A.: Getting to know your corpus. In: Text, Speech and Dialogue, Springer (2012) 3–15
8. Pomikálek, J., Rychlý, P., Kilgarriff, A.: Scaling to billion-plus word corpora. Volume 41. (2009) 3–13
9. Šmerk, P., Rychlý, P.: Majka – rychlý morfologický analyzátor. Technical report, Masarykova univerzita (2009)
10. Ústav Českého národního korpusu FF UK, Praha: Czech national corpus – SYN2000. Online: http://www.korpus.cz (2000)
11. Kovář, V., Horák, A., Jakubíšek, M.: Syntactic analysis using finite patterns: A new parsing system for czech. In: Human Language Technology. Challenges for Computer Science and Linguistics, Springer (2011) 161–171
12. Řehůřek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks. (2010) 46–50
13. Baisa, V., Suchomel, V.: Large corpora for turkic languages and unsupervised morphological analysis. In: Proceedings of LREC 2012 Workshop on Language Resources and Technologies for Turkic Languages, pp. 28–32. Istanbul, Turkey, 2012.

[2] `http://nlp.fi.muni.cz/projekty/corpora_comparison`