Recent Czech Web Corpora

Vít Suchomel

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
xsuchom2@fi.muni.cz

Abstract. This article introduces the largest Czech text corpus for language research – czTenTen12 with 5.4 billion tokens. A brief comparison with other recent Czech corpora follows.

Key words: web corpora, Czech

1 Introduction

Algorithms in the field of natural language processing generally benefit from large language models. Many words and phrases occur rarely, therefore there is a need for very large text collections to research behaviour of words. [1] Furthermore, the quality of the data obtained from the web is also stressed. [2] Language scientists are increasingly turning to the web as a source of language data. [3] Nowadays, the web is the biggest, easily exploitable and the cheapest source of text data.

We decided to support corpora based research of Czech language by building a new Czech corpus from web documents. The aim was to apply successful data cleaning tools and label the words with grammatical categories.

2 Building a new Czech web corpus

CzTenTen12 is a new Czech web corpus built in 2012 using data obtained from the web in 2011. Several automatic cleaning and postprocessing techniques were applied to the raw data to achieve a good quality corpus for language research.

2.1 Crawling the web

We used web crawler SpiderLing1, our previous work [4], to gather the data from the web. We started the crawl from 20000 seed URLs spanning over 8600 domains. The URLs were chosen using Corpus Factory [5], Czech Wikipedia and partially from older web corpus czTenTen. The crawl was restricted to the Czech national top level domain (.cz). 15 million documents of size 500 GB were downloaded in 24 days.

1http://nlp.fi.muni.cz/trac/spiderling
2.2 Postprocessing and tagging

The crawler performed character encoding detection and converted the data to UTF-8. The crawler detected language using character trigrams and filtered out texts in other languages than the focus language. Extra care had to be taken in case of Slovak which contains similar trigrams and unwanted texts may pass the filter. We prepared a list of Czech words not present in Slovak and a dual list of Slovak words not present in Czech. Using these lists, paragraphs containing three times more unique Slovak words than unique Czech words (0.4%) or unique words mixed from both languages (6.4%) were separated.

Boilerplate removal tool jusText\(^2\) was used to remove html markup, page navigation, very short paragraphs and other useless web content. The data was de-duplicated by removing exact and near duplicate paragraphs using tool onion\(^3\). Paragraphs containing more than 50% seen 7-grams were dropped.

Paragraphs containing only words without diacritical marks were tagged for further use, e.g. when studying the informal language of the web.\(^4\) Parts containing more than 20% of words not recognized by morphological analyzer Desamb were considered nonsense and removed from the corpus. The final size of the corpus reaches 5.4 billion tokens (4.4 billion words).

Czech morphological analyzer Desamb \[^{\text{\[6,7]\)}}\] was used to tag the corpus. The added information consists in the part of speech and other grammatical categories (where applicable): gender, number, case, aspect, modality and other.\(^5\)

3 Comparison with other corpora

The following recent Czech corpora are used in the comparison:

- SYN 2010 ...Czech national corpus – the SYN-series corpora up to 2010\(^6\),
- czes2 (a web corpus from 2009),
- czTenTen (a web corpus from 2011),
- the Hector project corpus\(^8\) (Hector) \[^2\].

3.1 Basic properties

According to \[^2\], both SYN and Hector are deliberately balanced, e.g. the latter consists of 450 millions of words from news and magazines, 1 billion of words

\(^2\)\text{http://nlp.fi.muni.cz/projects/justext}
\(^3\)\text{http://nlp.fi.muni.cz/projects/onion}
\(^4\) These texts come mostly from discussions or other informal websites (where people do not bother writing proper accent marks).
\(^5\) Reference of full tagset:\text{http://nlp.fi.muni.cz/projekty/ajka/tags.pdf}
\(^6\)\text{http://ucnk.ff.cuni.cz/english}
\(^7\) Although the full corpus is not publicly available, a wordlist with frequencies was enough to carry out measurements presented later on.
\(^8\)\text{http://hector.ms.mff.cuni.cz}
Table 1. Basic comparison of corpora. Only words consisting of letters are accounted in the word count. Dictionary size is the number of unique words with at least 5 occurrences. The the-score is the rank of word "the" in a list of words sorted by frequency from the most frequent one. The lower the value, the higher contamination by foreign words should be expected.

<table>
<thead>
<tr>
<th>corpus</th>
<th>word count \times 10^6</th>
<th>dictionary size \times 10^3</th>
<th>the-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>SYN2010</td>
<td>1300</td>
<td>1.61</td>
<td>7896</td>
</tr>
<tr>
<td>czes2</td>
<td>367</td>
<td>1.03</td>
<td>42</td>
</tr>
<tr>
<td>czTenTen</td>
<td>1652</td>
<td>2.42</td>
<td>1023</td>
</tr>
<tr>
<td>Hector</td>
<td>2650</td>
<td>2.81</td>
<td>1184</td>
</tr>
<tr>
<td>czTenTen12</td>
<td>4439</td>
<td>4.16</td>
<td>1223</td>
</tr>
</tbody>
</table>

Table 2. Corpus distance measured for each couple of corpora. The lower the distance score, the more similar is the couple.

<table>
<thead>
<tr>
<th>corpus</th>
<th>czes</th>
<th>czTenTen</th>
<th>Hector</th>
<th>czTenTen12</th>
</tr>
</thead>
<tbody>
<tr>
<td>SYN2010</td>
<td>1.60</td>
<td>1.70</td>
<td>2.28</td>
<td>1.73</td>
</tr>
<tr>
<td>czes2</td>
<td>1.44</td>
<td>1.79</td>
<td>1.52</td>
<td>1.12</td>
</tr>
<tr>
<td>czTenTen</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hector</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

from blogs and 1.1 billion of words from discussions. The content of the new corpus was not controlled and a deeper analysis of content remains for further research.

Table 1 displays values of three metrics calculated for five corpora. We observe czTenTen12 is the largest corpus with the largest dictionary. The the-score is a very simple metric offering a basic idea about contamination of the corpus by foreign (English) words. We observe czes2 is the most polluted corpus and SYN2010 is the most clean corpus in this measurement.

3.2 Corpora similarity

Table 2 shows a corpus comparison cross-table. The distance score calculation is based on relative corpus frequencies of 500 most frequent words in all corpora. The full method is described in [8]. We observe czTenTen and czTenTen12 are very close. That can be explained by similar way of obtaining and processing the data and sharing a lot of documents. On the other hand, the balanced corpora are more distant.

Comparison of keywords (also based on the relative corpus frequency) in czTenTen12 most different from SYN2010 was published in [9]. We observe there are more discussions and blogs (informal words, verbs in 1st or 2nd person, pronouns, adverbs) and computer related words in the new unbalanced corpus. Comparing czTenTen12 to Hector, we find the difference in presence of informal words too. Top czTenTen12 related words in this comparison are quite formal: již, lze, oblasti, společnosti, zařízení, této, roce, zde, mohou, rámci, projektu, těchto,
Fig. 1. Word sketch for word příst in cze2. A part of grammatical relations is displayed. The number of hits of the word in the corpus is 5191.

They could belong to some project notes or contracts. The key words of the opposite direction are no, holky, jo, xD, D, blog, teda, taky, já, dneska, sem, jdu, máš, which leads to conclusion Hector contains more blogs and discussions (generally informal texts) than czTenTen12.

3.3 Word sketches – bigger is better

Figures 1 and 2 display word sketch for word příst in cze2 and czTenTen12 in SketchEngine⁹. As can be easily observed, the bigger corpus offers better words in relations with the head word. E.g. (příst, blaho) in relation has_obj7 (which stands for a verb with a noun in instrumental case) is a quite common collocation in Czech. That is well reflected in the sketch for czTenTen12 with 84 occurrences and the first place by saliency score in the relation table. However, the smaller corpus offers only 4 instances of this collocation. Relation has_obj4 (which stands for a verb with a noun in accusative case) in cze2 is very poor, while containing many well suiting words in the case of the bigger corpus: len,
Fig. 2. Word sketch for word *příst* in *czTenTen12*. A part of grammatical relations is displayed. The number of hits of the word in the corpus is 28276, that is 5 times more frequent than in the smaller corpus.
4 Conclusion and future work

This article introduced the largest Czech text corpus for language research. A basic comparison with other contemporary Czech corpora was made. Example work sketches were shown to support idea that bigger corpora are better. The future plans for building web corpora of Slavonic languages include gathering resources in Polish and Croatian. Another interesting research opportunity is studying semantic topics automatically extracted from documents in the corpus. That would help us to know more about the content of the corpus and consequently of the Czech web.

Acknowledgements

This work has been partially supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013, by the Ministry of the Interior of CR within the project VF20102014003 and by the Czech Science Foundation under the project P401/10/0792.

References


10 Presence of other words in this relation is caused by tagging mistakes or by putting them in a wrong relation.