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 colletions 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 *SpiderLing*[1], 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.

[1] `http://nlp.fi.muni.cz/trac/spiderling`

Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2012, pp. 77–83, 2012.

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* [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][7],
- *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] `http://nlp.fi.muni.cz/projects/justext`

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

[4] These texts come mostly from discussions or other informal websites (where poeople do not bother writing proper accent marks).

[5] Reference of full tagset: `http://nlp.fi.muni.cz/projekty/ajka/tags.pdf`

[6] `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] `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.

corpus	word count [10^6]	dictionary size [10^6]	the-score
SYN2010	1300	1.61	7896
czes2	367	1.03	42
czTenTen	1652	2.42	1023
Hector	2650	2.81	1184
czTenTen12	4439	4.16	1223

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

corpus	czes	czTenTen	Hector	czTenTen12
SYN2010	1.60	1.70	2.28	1.73
czes2		1.44	2.16	1.52
czTenTen			1.79	1.12
Hector				1.65

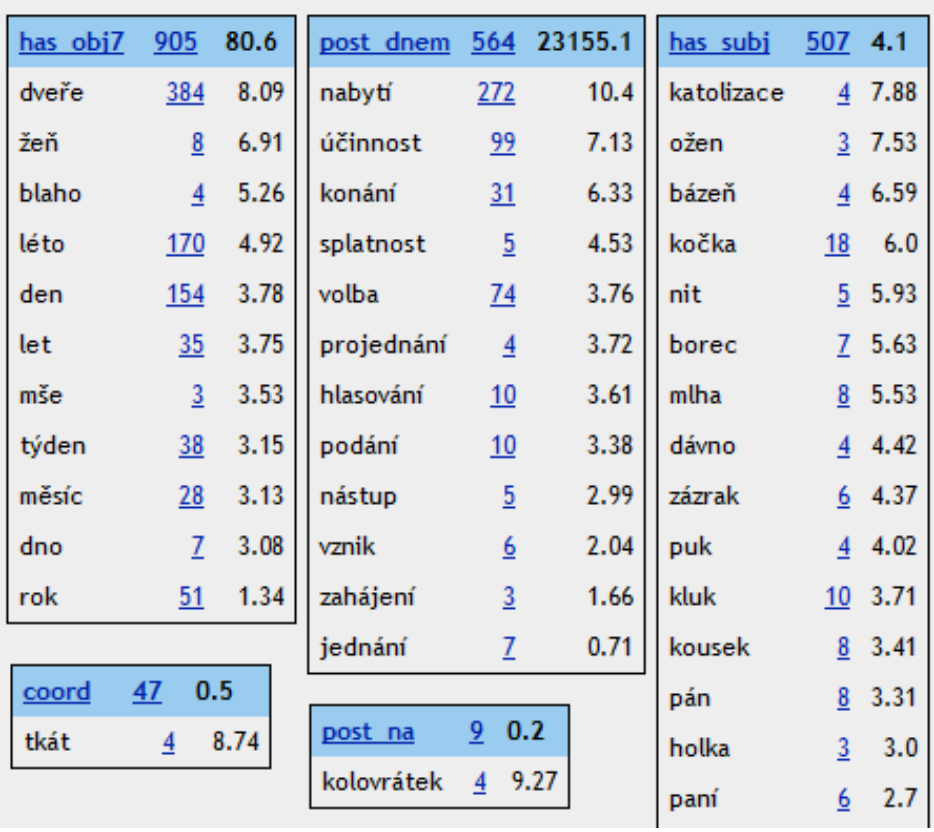
from blogs and 1.1 billion of words from discussions. The content of the new corpus was not controlled and a deeper analysis of content remains for further research.

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 differrence 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*,



has_obj7	905	80.6
dveře	384	8.09
žeň	8	6.91
blaho	4	5.26
léto	170	4.92
den	154	3.78
let	35	3.75
mše	3	3.53
týden	38	3.15
měsíc	28	3.13
dno	7	3.08
rok	51	1.34

post_dnem	564	23155.1
nabytí	272	10.4
účinnost	99	7.13
konání	31	6.33
splatnost	5	4.53
volba	74	3.76
projednání	4	3.72
hlasování	10	3.61
podání	10	3.38
nástup	5	2.99
vznik	6	2.04
zahájení	3	1.66
jednání	7	0.71

has_subj	507	4.1
katolizace	4	7.88
ožen	3	7.53
bázeň	4	6.59
kočka	18	6.0
nit	5	5.93
borec	7	5.63
mlha	8	5.53
dávno	4	4.42
zázrak	6	4.37
puk	4	4.02
kluk	10	3.71
kousek	8	3.41
pán	8	3.31
holka	3	3.0
paní	6	2.7

coord	47	0.5
tkát	4	8.74

post_na	9	0.2
kolovrátek	4	9.27

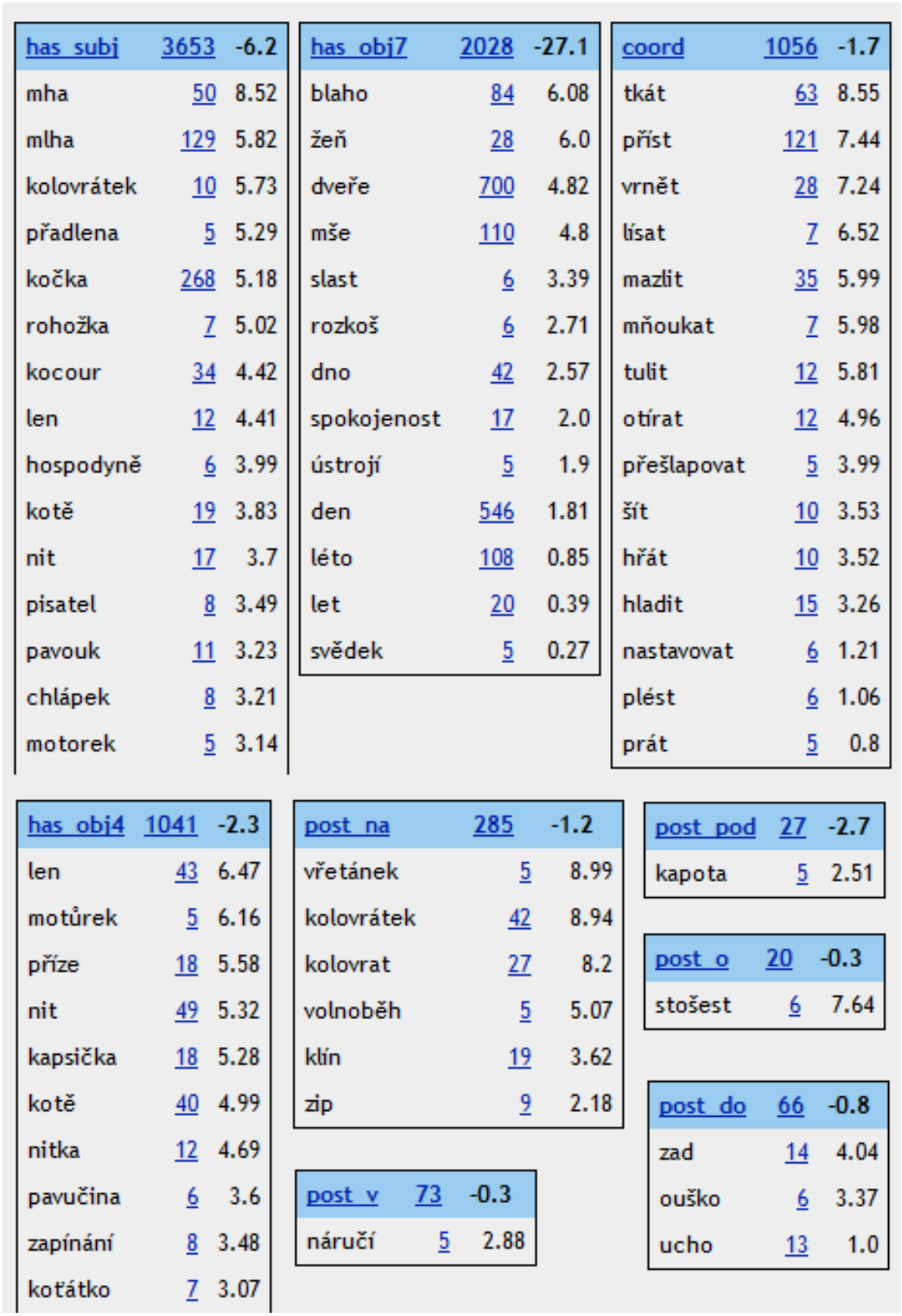
Fig. 1. Word sketch for word *přist* in *czes2*. A part of grammatical relations is displayed. The number of hits of the word in the corpus is 5191.

systému. 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 *přist* in *czes2* and *czTenTen12* in SketchEngine[9]. As can be easily observed, the bigger corpus offers better words in relations with the head word. E.g. (*přist*, *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 *czes2* is very poor, while containing many well suiting words in the case of the bigger corpus: *len*,

[9] `http://sketchengine.co.uk/`



has_subj	3653	-6.2
mha	50	8.52
mlha	129	5.82
kolovrátek	10	5.73
přadlena	5	5.29
kočka	268	5.18
rohožka	7	5.02
kocour	34	4.42
len	12	4.41
hospodyně	6	3.99
kotě	19	3.83
nit	17	3.7
pisatel	8	3.49
pavouk	11	3.23
chlápek	8	3.21
motorek	5	3.14

has_obj7	2028	-27.1
blaho	84	6.08
žeň	28	6.0
dveře	700	4.82
mše	110	4.8
slast	6	3.39
rozkoš	6	2.71
dno	42	2.57
spokojenost	17	2.0
ústrojí	5	1.9
den	546	1.81
léto	108	0.85
let	20	0.39
svědek	5	0.27

coord	1056	-1.7
tkát	63	8.55
přísť	121	7.44
vrnět	28	7.24
lísat	7	6.52
mazlit	35	5.99
mňoukat	7	5.98
tulit	12	5.81
otírat	12	4.96
přešlapovat	5	3.99
šít	10	3.53
hřát	10	3.52
hladit	15	3.26
nastavovat	6	1.21
plést	6	1.06
prát	5	0.8

has_obj4	1041	-2.3
len	43	6.47
motůrek	5	6.16
příze	18	5.58
nit	49	5.32
kapsička	18	5.28
kotě	40	4.99
nitka	12	4.69
pavučina	6	3.6
zapínání	8	3.48
koťátko	7	3.07

post_na	285	-1.2
vřetánek	5	8.99
kolovrátek	42	8.94
kolovrat	27	8.2
volnoběh	5	5.07
klín	19	3.62
zip	9	2.18

post_v	73	-0.3
náručí	5	2.88

post_pod	27	-2.7
kapota	5	2.51

post_o	20	-0.3
stošest	6	7.64

post_do	66	-0.8
zad	14	4.04
ouško	6	3.37
ucho	13	1.0

Fig. 2. Word sketch for word *přísť* 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.

příze, *nit*, *nitka*, *pavučina*.[10] Furthermore, most of collocations with prepositions in prepositional relations *post_na*, *post_v*, *post_do* and other are present just in the word sketch of the bigger corpus: *na kolovratu*, *na klíně*, *v náručí*, *do ouška*, *pod kapotou*, *ostošest*. We conclude a bigger corpus is much more useful for language research based on collocations of words.

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

1. Pomikálek, J., Rychlý, P., Kilgarriff, A.: Scaling to billion-plus word corpora. Advances in Computational Linguistics **41** (2009) 3–13
2. Spoustová, J., Spousta, M.: A high-quality web corpus of czech. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, European Language Resources Association (ELRA) (may 2012)
3. Kilgarriff, A., Grefenstette, G.: Introduction to the special issue on the web as corpus. Computational linguistics **29**(3) (2003) 333–347
4. Suchomel, V., Pomikálek, J.: Efficient web crawling for large text corpora. In: Proceedings of the Seventh Web as Corpus Workshop, Lyon, France (2012)
5. Kilgarriff, A., Reddy, S., Pomikálek, J., Pvs, A.: A corpus factory for many languages. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'10, Malta) (2010)
6. Šmerk, P.: Unsupervised Learning of Rules for Morphological Disambiguation. In: Lecture Notes in Artificial Intelligence 3206, Proceedings of Text, Speech and Dialogue 2004, Berlin, Springer-Verlag (2004) 211–216
7. Jakubíček, M., Horák, A., Kovář, V.: Mining phrases from syntactic analysis. In: Lecture Notes in Artificial Intelligence, Proceedings of Text, Speech and Dialogue 2009, Plzeň, Czech Republic, Springer-Verlag (2009) 124–130

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

8. Kilgarriff, A.: Comparing corpora. International journal of corpus linguistics **6**(1) (2001) 97–133
9. Kilgarriff, A.: Getting to know your corpus. In: Text, Speech and Dialogue, Springer (2012) 3–15