csTenTen17, a Recent Czech Web Corpus

Vít Suchomel[1,2]

[1] Lexical Computing
[2] NLP Centre, Masaryk University, Brno
xsuchom2@fi.muni.cz

Abstract. This article introduces a very large Czech text corpus for language research – *csTenTen17* compiled from texts downloaded in 2015, 2016 and 2017. The corpus is consisting of 10.5 billion words reaching double the size of its predecessor from 2012. A brief comparison with other recent Czech corpora follows.

Keywords: Czech corpus; web corpus; text processing

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 the behaviour of words. [10]. Furthermore, the quality of the data obtained from the web is also important. [13] Linguists studying natural languages, lexicographers compiling dictionaries, sociologists studying the topics moving the society, marketing experts creating brand names, language engineers building language models and many others are turning to the web as a source of language data. 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 again by building an up-to-date corpus from web documents in Czech. The aim was to apply text cleaning software, language discrimination tools, and deduplication to a corpus of a ten billion words size. The corpus should be indexed in a corpus manager providing a basic concordance search as well as advanced functions such as a summary of grammatical and collocational behaviour of words.

1.1 Paper Outline

Corpus construction and properties are described in Section 2. The result corpus is compared to other Czech corpora in Section 3. Final remarks are presented in Section 4.

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2018, pp. 111–123, 2018.

2 Corpus Construction And Properties

2.1 Crawling The Czech Web

The corpus consists of texts obtained using crawler SpiderLing [14]. The crawler collected texts from the web in October and November 2015, October and November 2016, and May, October and November 2017. The crawler started from seed web domains and URLs coming from various sources:

- csTenTen12 document sources (the previous Czech web corpus),
- lists of web domains presenting a good quality content such as `dmoz.org` and `urlblacklist.com`[3],
- URLs of Czech documents obtained by querying search engine Bing for Czech words,
- manually selected Czech web news sites (`blisty.cz`, `ihned.cz`, `lidovky.cz`, `novinky.cz`, `reflex.cz`, `seznam.cz`).

The crawler was not restricted to download just from the Czech national top level domain `.cz`. It was set not to crawl web sites not providing Czech text and to slightly prefer web sites yielding more Czech text than other domains.

Data processing tasks important for the crawler to evaluate the yield rate of Czech text of web sites were carried out by tools embedded within the crawler:

- Encoding detection using byte trigram models by Chared[4] [11],
- language identification on the document level using character trigram models,
- HTML boilerplate removal by Justext 1.4[5] [9],
- splitting text to paragraphs by Justext using HTML tags `<p>`, `<div>` and `<br>`,
- language checking on the paragraph level using lists of frequent words by Justext,
- exact duplicate removal on the document level using hashes of HTML data and plain text.

All models necessary for the process were built using samples of Czech text or web pages from the Czech web before starting the crawler.

The following sizes apply just to the 2017 batch: The crawler made 590 million HTTP requests to internet servers. 7.0 TB of raw HTTP response data containing 150 million web pages were collected. Of these, 35 million web pages contained at least one paragraph of Czech text recognised by Justext[6]. The size of the plain text obtained by the crawler before additional filtering described in Section 2.3 was 60 GB.

[3] Both of these web domain catalogues are no longer available on the web in 2018.

[4] `http://corpus.tools/wiki/Chared`

[5] `http://corpus.tools/wiki/Justext`

[6] The following Justext parameters were used to recognise paragraphs of text long enough: `length_low = 70`, `length_high = 140` (200 by default), `stopwords_low = 0.2` (0.3 by default), `stopwords_high = 0.3` (0.32 by default), `max_link_density = 0.4` (0.2 by default). The default values were altered to allow slightly shorter paragraphs to extract more text while keeping the level of strictness high.

2.2 Collecting Texts From Wikipedia

Since Wikipedia pages share the structure of a document and they are coded in MediaWiki markup language[7] which is not straightforward to turn into a plain text, software Wiki2Corpus[8] was used to obtain texts from the Czech Wikipedia for the corpus. The tool ran in November 2017 and aimed for both encyclopedia articles and respective talk pages[9].

642,693 Wikipedia pages or 15 GB of data were downloaded by Wiki2Corpus (including talks, redirections, disambiguation pages) and converted from the MediaWiki markup to documents consisting of plain text. The size of the data after extracting paragraphs using Justext was 1.0 GB. Most stubs or other short articles were discarded because they were lacking nice long paragraphs recognised by Justext.

The plain text was further cleaned and filtered using the same methods as the crawled web pages. The process is described in the following section.

2.3 Postprocessing of Text

Methods of postprocessing of corpus plain text after the crawling applied to all parts of the corpus are described in this section. The sizes however represent only the part of data collected in 2017 since the information about processing the parts from 2015 and 2016 are no longer available.

The plain text was split to tokens using Unitok [15]. The size of the 2017 data at this stage of processing was 7.8 billion tokens.

Despite the character n-gram model based removal of documents in other than the target language, there were still a lot of paragraphs in unwanted languages (i.e. other than Czech, especially English and Slovak). Language seperation based on a method exploiting large lists of word forms with relative corpus frequency[10] in large monolingual web corpora[11] described in [3] was applied to paragraphs and documents of the tokenised text.

Czech, Czech without diacritics, Slovak, Slovak without diacritics, English, German, Polish, Slovene, Croatian, Russian, French, Spanish, and Italian were discerned. Only the Czech part (with diacritic marks) was allowed to get to the final corpus. 0.1 % of paragraphs were filtered out because the majority of the content was not in Czech, 1.0 % of paragraphs were thrown away because of a content in multiple languages, and 3.6 % of paragraphs were filtered out since they were too small to reliably determine a language (in fact, these paragraphs

[7] `https://www.mediawiki.org/wiki/Markup_spec`

[8] `http://corpus.tools/wiki/wiki2corpus`

[9] E.g. `https://cs.wikipedia.org/wiki/1984_(roman)` and its talk page `https://cs.wikipedia.org/wiki/Talk:1984_(roman)`

[10] Relative corpus frequency is the number of occurrences of a word form per billion tokens in the corpus.

[11] Web corpora built in the past were used. In case there was no corpus in the target language, the list would ben obtained by bootstrapping, i.e. applying the same method several times to the corpus until the result frequency list stops changing.

would not contribute much to the quality of the corpus even though they were in Czech). 4.7 % of paragraphs were removed in total in this step of postprocessing the data.

Examples of paragraphs of text removed because of the relative frequency of word forms is larger in a reference web corpus of other language than in csTenTen12 follow (non-Czech words are striked out). Example 1[12]: *23 ~~Solo Pieces for La Naissance de L'Amour~~ je soundtracková album velšského multiinstrumentalisty Johna Calea. Album vyšlo v roce 1993 u vydavatelství ~~Les Disques du Crépuscule~~. Album produkoval ~~Jean-Michel Reusser~~.* Example 2[13]: *~~Nice~~ hotel ~~at~~ a ~~good location. Rooms very good, but beds~~ a ~~little bit hard. The staff was nice and helpful. Nice location close~~ to ~~Konakli center with lot of shops and market~~ on ~~Wednesdays. Nice~~... celá recenze s možností překladu.*

Near-duplicate paragraph deduplication was carried out using Onion[14] [9], a tool based on comparing hashes of n-grams of tokens. In the case of this corpus, paragraphs containing more than 90 % of 5-tuples of tokens seen before (i.e. in a part of the input read earlier) were removed. The smoothing mode was on with the minimum length of a stub set reduced to 10 tokens.[15]

The text was split to sentences using a tool looking for fullstops (or other end of sentence markers) followed by a space and a capital letter and dealing with abbreviations according to a predefined list.

2.4 Morphological Annotation

The corpus was lemmatised and morphologically annotated using Czech morphological analyzer Majka [17]. The analyser determined the part of speech and other grammatical categories (where applicable): gender, number, case, aspect, modality and other.[16] The tags were desambiguated by Desamb [12,4]. A gender respecting lemma was added to allow creating name phrases from lemmas properly.[17]

The most frequent parts of speech identified in the corpus are nouns (33 %), verbs (16 %), adjectives (12 %), prepositions (10 %), pronouns (9 %), and adverbs (7 %).[18]

[12] Text source: `https://cs.wikipedia.org/wiki/23_Solo_Pieces_for_La_Naissance_de_L'Amour`

[13] Text source: `https://www.ellagris.cz/turecko/turecka-riviera/alanya/royal-garden-select-626634`

[14] http://corpus.tools/wiki/Onion

[15] The full parameters: `onion -s -n 5 -t 0.9 -l 10`. More about tuning the parameters of onion can be found in a paper by V. Benko [1].

[16] See `https://www.sketchengine.co.uk/tagset-reference-for-czech` for the full tagset reference.

[17] For example, the base form of "veřejné knihovně" is not "veřejný knihovna" where "veřejný" (masculine) is the lemma of "veřejná" (feminine) but "veřejná knihovna" where the gender of the noun is respected by the adjective properly.

[18] Not counting the punctuation, abbreviations, foreign words.

Table 1: Sizes of parts of csTenTen17 by the source subcorpus

	Total	Wiki 17	Wiki Talk	Web 17	Web 16	Web 15
Tokens	12,586,415,546	0.97 %	0.09 %	58 %	32 %	8.4 %
Words	10,502,222,474					
Sentences	738,085,256					
Paragraphs	227,097,470					
Documents	35,995,251	1.00 %	0.09 %	62 %	30 %	7.6 %

Table 2: Lexicon sizes

Word form	Lemma	Gender respecting lemma	Tag	Part of speech
40,445,706	29,100,249	34,014,060	2,247	15

2.5 Final Sizes

The final corpus consists of 12,586,415,546 tokens in 35,995,251 documents. 91 % of tokens of the final corpus come from the Czech TLD .cz. Sizes of parts of the corpus by the source can be found in Table 1. Sizes of lexicons are in Table 2. Document counts in TLDs and web sites are presented by Table 3.

Table 3: Document count – the largest web domains and domain size distribution

TLDs		Web domains		Web domain size distribution	
cz	91 %	webnode.cz	660,000	At least 1 document	350,000
com	2.3 %	idnes.cz	540,000	At least 5 documents	190,000
eu	1.9 %	blogspot.cz	450,000	At least 10 documents	130,000
org	1.8 %	wikipedia.org	390,000	At least 50 documents	53,000
net	1.2 %	lidovky.cz	180,000	At least 100 documents	34,000
info	1.0 %	zive.cz	170,000	At least 500 documents	9,100
		tyden.cz	150,000	At least 1,000 documents	4,900
		estranky.cz	130,000	At least 5,000 documents	950
		e15.cz	120,000	At least 10,000 documents	460
		denik.cz	120,000	At least 50,000 documents	38
		tiscali.cz	120,000	At least 100,000 documents	13
		sluzby.cz	110,000	At least 500,000 documents	2
		rozhlas.cz	110,000		
		mobilmania.cz	100,000		
		penize.cz	98,000		
		ihned.cz	97,000		

2.6 Access To The Corpus

Since the corpus is a part of the HaBiT project[19] [8], it can be accessed via corpus manager Sketch Engine [6] at the project site.[20]. Functionality provided by Sketch Engine covers concordance search, wordlist search, collocation and word frequency calculation, Word Sketches, thesaurus and more.

3 Comparison With Other Recent Corpora

Our older paper on a Czech web corpus from 2012 is followed in this section. [16] There are the following recent Czech corpora used in the comparison:

- csTenTen17 – the new corpus,
- czTenTen12 (v. 9) – the previous version of csTenTen from 2012,
- Araneum Bohemicum III Maius (17.04, v. 1.3.61) – web corpus downloaded by V. Benko from 2013 to 2016. Crawled and processed by similar tools as in the case of TenTen corpora. [2]
- csSkELL (v. 2.2) – Czech web corpus of example sentences gained from websites provided by Czech WebArchive to 2016. Processed by similar tools as TenTen corpora.[21]
- SYN 2015 – Czech national corpus, a reference representative corpus containing fiction, non-fiction and journalism texts mostly from 2010 to 2014.[22] [7] This corpus is a non-web ballanced and representative corpus to compare less controlled web corpora to.[23]

3.1 Basic Properties

Tables 4 and 5 display values of six metrics calculated for the compared corpora. We observe the largest corpus has the largest dictionary and the least varied vocbulary.

Documents in csTenTen17 are shorter than in its predecessor. That might be caused by a similar composition of genres in the web, e.g. not much fiction that tends to contain long documents. The length of sentences is quite similar for all selected corpora.

csTenTen17 may be the corpus least contaminated by foreign text. That can be explained by an additional method of removing unwanted languages described in Section 2.3.

[19] `https://habit-project.eu/`

[20] `https://corpora.fi.muni.cz/habit/run.cgi/first?corpname=cstenten17_mj2`

[21] `https://www.sketchengine.co.uk/cskell/`

[22] `https://www.korpus.cz/`

[23] Although the full text of the corpus is not publicly available, a wordlist with frequencies was enough to carry out wordlist based measurements.

Table 4: Basic comparison of corpora: Token counts and type-token ratio. The higher TTR, the more varied vocabulary.

Corpus	Token count	Word lexicon	Type-token ratio
csTenTen17	12,600,000,000	40,400,000	0.003,2
czTenTen12	5,070,000,000	18,700,000	0.003,7
Araneum Bohemicum	1,200,000,000	8,460,000	0.007,1
csSkELL	1,730,000,000	8,010,000	0.004,6

Table 5: Basic comparison of corpora: Average document length (the number of tokens) (structure `<text>` used in the case of SYN 2015), average sentence length, "the-score". The-score, being the rank of word "the" in a list of lowercased words, is a very simple metric offering a basic idea about contamination of the corpus by foreign (English) text. The higher the value, the better.

Corpus	Avg. doc tokens	Avg. sentence tokens	The-score
csTenTen17	350	17	7,387
czTenTen12	550	18	730
Araneum Bohemicum	460	17	517
csSkELL	N/A	19	475
SYN 2015	1,055	15	1,145

3.2 Keyword Comparison

Keyword comparison as a way of telling differences between corpora was performed by Kilgarriff in [5]. Using the same method – putting csTenTen17 as the focus corpus and other corpora in the place of the reference corpus – the words with the highest relative frequency in comparison to words in other corpus or subcorpus are the highest ranked by the keyword score:

$$keywordscore = \frac{fpm_{foc}(w) + n}{fpm_{ref}(w) + n}$$

where $fpm(w)$ represents occurrences per million of word w, foc is the focus corpus, ref is the reference corpus, and n is a smoothing parameter.

Table 6 shows differences in the content of csTenTen17 in comparison to other corpora. It can be observed the new corpus covers topics trending recently such as "babiš", "eet", "trump", "sýrii", "krymu", "instagram", "severus", "snape", "naruto", "parlamentnílisty". (The last might be a tokenisation error as well.) There is also a lot of finance and trade related material in the 2017 corpus, e.g. "půjčka", "půjčky", "nebankovní", "směnnost", "prodám", "skladem". These words may indicate the presence of non-text in the corpus that should be investigated (short phrases without subject predicate pairs, or even computer

Table 6: Keyword comparison of csTenTen17, the 2017 subcorpus, to other corpora. Settings: lowercased word forms, minimum frequency 10, smoothing parameter 1 preferring rare words over common words.

	csTenTen12		Araneum Bohemicum		csSkELL	
Rank	**Word**	**Score**	**Word**	**Score**	**Word**	**Score**
1	pujcka	16.9	odst	56.8	č	49.4
2	babiš	14.2	písm	21.3	půjčka	15.4
3	pujcky	13.3	č	21.0	vč	14.4
4	půjčka	9.6	vč	12.3	prodám	13.8
5	trump	9.2	hellip	8.2	pujcka	13.6
6	babiše	8.7	tis	8.2	pujcky	10.9
7	eet	8.2	zák	7.7	nbsp	8.1
8	č	7.4	obr	7.5	hellip	7.7
9	trumpa	6.3	mld	6.9	naruto	7.7
10	azure	6.2	hl	6.4	kdyz	7.6
11	zadavatel	5.2	ust	5.7	ú	7.1
12	ú	4.9	okr	5.6	skladem	6.9
13	gmbh	4.8	naruto	4.9	nabízíme	6.9
14	sýrii	4.6	atp	4.8	severus	6.7
15	dodavatelský	4.4	ú	4.5	panička	6.5
16	zadavatele	4.4	parlamentnílisty	4.4	snape	6.4
17	nebankovní	4.3	azure	4.3	nebankovní	6.3
18	půjčky	4.1	dodavatelský	4.1	kontaktujte	6.2
19	krymu	4.1	protoe	4.0	koupelna	6.0
20	směnnost	4.1	oponent	4.0	pleť	5.8
21	instagram	4.1	xvi	3.9	půjčky	5.7
22	vyžádejte	4.0	ev	3.7	pred	5.7

generated text). There is also a lot of differences in tokenisation, especially in the case of Araneum: “odst”, “obr”, “vč” (these abbreviations were not recognised in our corpus). Word “protoe” may be a misspelling. Finally, some words without diacitics scored high, e.g. “pujcka”, “kdyz”, “pred”.

Keyword comparison to a non-web representative ballanced corpus shown in Table 7 reveals the new corpus contains relatively a lot of money lending text and also some internet related technical words.

Table 8 shows the 2016 subcorpus is polluted with an online gambling related spam.

3.3 Word Sketches

Multi Word Sketch for “chytat stéblo” (“to grasp a straw”, usually found in idiom “tonoucí se stébla chytá” – “grasping at straws”) in csTenTen17 and czTenTen12 are displayed in screenshots from Sketch Engine in Figures 1 and 2. As can be

Table 7: Keyword comparison of csTenTen17, the 2017 subcorpus, to SYN 2015 with the same settings as in Table 6. Lines affected by a different tokenisation, or misspellings were omitted to focus on differences in text type and genre.

	SYN 2015	
Rank	**Word**	**Score**
3	půjčka	27.1
7	prodám	17.3
8	nabízíme	16.5
11	nebankovní	15.4
13	klikněte	14.4
16	kontaktujte	13.0
20	skladem	11.6
21	naruto	11.6
23	půjčky	10.8
26	email	9.7
27	ikdyž	9.6
28	neváhejte	9.4
29	zadavatel	9.4
30	php	9.3
31	html	9.2
32	ahojky	9.2
33	online	9.2
35	trump	9.1

Table 8: Keyword comparison of the 2016 subcorpus to the 2017 part of csTenTen17 with the same settings as in Table 6.

	csTenTen17	
Rank	**Word**	**Score**
1	kasino	1479
2	sloty	1299
3	casino	969
4	automaty	708
5	kasina	649
6	kasinu	471
7	kasinové	463
8	hazardní	443
9	sajid	432
10	blackjack	422
11	jackpoty	408
12	jackpot	400
13	roztočení	385
14	lisu	309
15	beste	305
16	slot	301
17	činohra	301
18	zelenom	294
19	sizzling	267
20	karolínka	246

seen, the bigger corpus provides more collocations to study the meaning of the phrase. For example, "chytat stéblo záchrany" ("to grasp a straw of rescue") can be found only in a single case in the 2012 version of the corpus while there are four occurrences of the phrase in the new data.

"Chytat stéblo" is located in csSkELL, the smallest corpus in the comparison, only five times which is not enough to get relevant information about the phrase. Word Sketches of Araneum Bohemicum are not compared since the corpus is tagged by another tagger and its Word Sketches are based on a different grammar.

3.4 Thesaurus

According to our inspection of a computer generated thesaurus based on words sharing the same collocations in relations in Word Sketches, the size of a corpus contributes to finding better synonym candidates for low

Fig. 1: Multi Word Sketch for "chytat stéblo" ("to grasp a straw") in csTenTen17. Collocations occurring at least three times are displayed in several grammatical relations. The number of "chytat" collocating with "stéblo" in the corpus is 903.

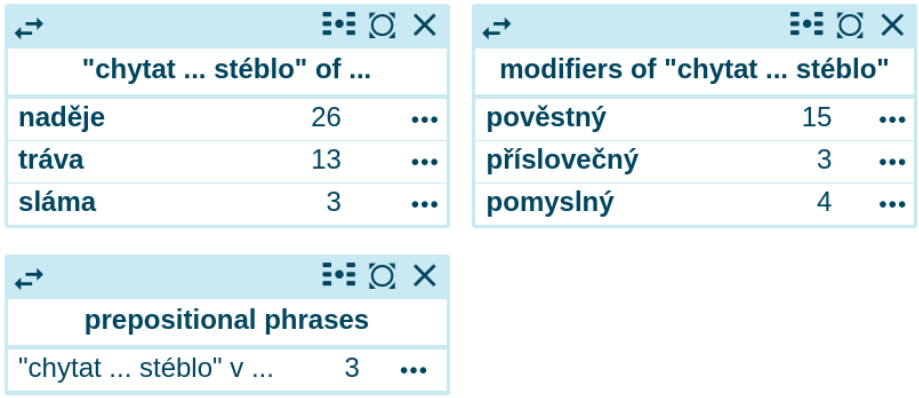

Fig. 2: Multi Word Sketch for "chytat stéblo" ("to grasp a straw") in csTenTen12. Collocations occurring at least three times are displayed in several grammatical relations. The number of "chytat" collocating with "stéblo" in the corpus is 506.

	CsTenTen17	Frequency	Cluster
1	prehistorický	17,885	pravěký 39,218
2	obstarožní	5,849	
3	rozhrkaný	1,235	otřískaný 2,208
4	lidožravý	2,248	
5	humanoidní	5,611	
6	ohyzdný	4,195	šeredný 4,196

	CzTenTen12	Frequency	Cluster
1	prehistorický	8,754	pravěký 18,423
2	humanoidní	2,994	
3	lochneský	183	lochnesský 143 lochnesské 57
4	obstarožní	3,035	
5	porouchaný	4,325	
6	druhohorní	1,851	třetihorní 2,067

	CsSkELL	Frequency	Cluster
1	prehistorický	3,220	pravěký 7,574
2	druhohorní	1,003	třetihorní 1,196
3	vyhynulý	2,256	vymřelý 1,047
4	potopní	67	
5	mýtický	3,216	mytický 2,558 mytologický 3,453
6	lochneský	34	Lochnesský 61 Lochneský 25

Fig. 3: Thesaurus of word “předpotopní” (“antediluvian”, “prehistoric”) based on words sharing the same collocations in relations in Word Sketches in three corpora. Note both csTenTen17 and czTenTen12 provide candidates meaning old, battered, chipped (“obstarožní”, “rozhrkaný”, “otřískaný”) while these important synonyms were not extracted from smaller csSkELL.

frequency words. For example, there are better results for adjective "předpotopní" ("antediluvian", "prehistoric") extracted from csTenTen17 (12 bn. tokens) and czTenTen12 (5.1 bn. tokens) than from csSkELL (1.7 bn. tokens) as can be seen on Figure 3.

4 Conclusion and Future Work

A new ten-billion-word Czech corpus was built from documents recently published on the web. The corpus can be searched by a publicly accessible corpus manager.

To focus on quality of the data, which is important for all kinds of corpus use, we would like to correct the tokenisation of abbreviations and to address the part of the corpus from 2016 containing online gambling advertisement spam. Furthermore, the users of the corpus would benefit from identification of topics and genres of documents. That will be another field to focus on in the future.

Acknowledgements. This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin infrastructure LM2015071.

References

1. Benko, V.: Data deduplication in slovak corpora. Slovko 2013: Natural Language Processing, Corpus Linguistics, E-learning pp. 27–39 (2013)
2. Benko, V.: Aranea: Yet another family of (comparable) web corpora. In: International Conference on Text, Speech, and Dialogue. pp. 247–256. Springer (2014)
3. Herman, O., Suchomel, V., Baisa, V., Rychlý, P.: Dsl shared task 2016: Perfect is the enemy of good language discrimination through expectation-maximization and chunk-based language model. In: Nakov, P., Zampieri, M., Tan, L., Ljubešić, N., Tiedemann, J., Malmasi, S. (eds.) Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3). pp. 114–118. Association for Natural Language Processing (ANLP) (2016), https://aclanthology.info/pdf/W/W16/W16-4815.pdf
4. Jakubíček, M., Horák, A., Kovář, V.: Mining phrases from syntactic analysis. In: Lecture Notes in Artificial Intelligence, Proceedings of Text, Speech and Dialogue 2009. pp. 124–130. Springer-Verlag, Plzeň, Czech Republic (2009)
5. Kilgarriff, A.: Getting to know your corpus. In: Text, Speech and Dialogue. pp. 3–15. Springer (2012)
6. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V.: The sketch engine: ten years on. Lexicography **1** (2014), http://dx.doi.org/10.1007/s40607-014-0009-9
7. Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kováříková, D., Petkevič, V., Procházka, P., et al.: Syn2015: representative corpus of contemporary written czech. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016. pp. 2522–2528 (2016)
8. Pala, K., Horák, A., Rychlý, P., Suchomel, V., Baisa, V., Jakubíček, M., Kovář, V., Nevěřilová, Z., Rambousek, A., Gambäck, B., Sikdar, U., Bungum, L.: Habit system (2017), http://corpora.fi.muni.cz/habit/

9. Pomikálek, J.: Removing Boilerplate and Duplicate Content from Web Corpora. Ph.D. thesis, Masaryk University, Brno (2011)
10. Pomikálek, J., Rychlý, P., Kilgarriff, A.: Scaling to billion-plus word corpora. Advances in Computational Linguistics **41**, 3–13 (2009)
11. Pomikálek, J., Suchomel, V.: chared: Character encoding detection with a known language. In: Aleš Horák, P.R. (ed.) Fifth Workshop on Recent Advances in Slavonic Natural Language Processing. pp. 125–129. Tribun EU, Brno, Czech Republic (2011)
12. Šmerk, P.: Unsupervised Learning of Rules for Morphological Disambiguation. In: Lecture Notes in Artificial Intelligence 3206, Proceedings of Text, Speech and Dialogue 2004. pp. 211–216. Springer-Verlag, Berlin (2004)
13. 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). European Language Resources Association (ELRA), Istanbul, Turkey (may 2012)
14. Suchomel, V., Pomikálek, J.: Efficient web crawling for large text corpora. In: Proceedings of the Seventh Web as Corpus Workshop. Lyon, France (2012)
15. Suchomel, V., Michelfeit, J., Pomikálek, J.: Text tokenisation using unitok. In: Eighth Workshop on Recent Advances in Slavonic Natural Language Processing. pp. 71–75. Tribun EU, Brno (2014)
16. Suchomel, V.: Recent czech web corpora. In: Aleš Horák, P.R. (ed.) 6th Workshop on Recent Advances in Slavonic Natural Language Processing. pp. 77–83. Tribun EU (2012)
17. Šmerk, P., Rychlý, P.: Majka – rychlý morfologický analyzátor. Tech. rep., Masarykova univerzita (2009), `http://nlp.fi.muni.cz/ma/`