csTenTen17, a Recent Czech Web Corpus

Vít Suchomel

vit.suchomel@sketchengine.co.uk

December 8, 2018
Raslan 2018

Outline

Recent Czech Big Corpora

- czTenTen12 (2011)
- Hector (2011)
- SYN 2015 (2010–2014)
- Araneum Bohemicum (2013–2016)
- csSkELL (2016)
- csTenTen17 (2015–2017)

Why To Build New Corpora?

- Up-to-date written form of the target language
- Text from a particular year to add to a diachronic corpus
- The web grows $\rightarrow$ the corpus size grows

Crawling – Crawler SpiderLing

- Sites providing Czech text (no limit to the Czech TLD)
- Oct – Nov 2015, Oct – Nov 2016, and May, Oct, Nov 2017

Crawling – Seed URLs

- csTenTen12 document sources
- Good quality content site lists `dmoz.org` and `urlblacklist.com`
- Bing search results for Czech words
- Czech web news sites (`blisty.cz`, `ihned.cz`, `lidovky.cz`, `novinky.cz`, `reflex.cz`, `seznam.cz`

Data Processing Steps I

- Encoding detection (Chared = byte trigrams)
- Language identification
  - Documents (Justext = frequent Czech words)
  - Paragraphs (character trigrams)
  - Documents & paragraphs (wordlists of 13 languages)
- HTML splitting to paragraphs & boilerplate removal (Justext)
- Duplicate removal
  - Exact documents (text hashes)
  - Similar paragraphs (Onion = word 5-grams)

5 % of paragraphs were removed

Examples of removed text (non Czech words in bold):

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****.*[1]

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.*[2]

---

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

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

Corpus Frequency Wordlists (Sample)

Czech		Czech no diac.		Slovak		Slovak no diac.	
který	8395917	bylo	6895481	keď	1222751	ze	910030
bude	8367336	aby	6637007	ak	1155099	mi	882506
byl	8026382	byla	6411943	vo	1143786	ani	849737
mi	7559285	pak	6322052	ktorý	1040954	ho	793819
má	7497525	ze	5898536	jeho	1022330	bolo	789226
také	7288326	pokud	5842344	bol	926324	tu	771362
jeho	7279924	ani	5674096	bude	922604	i	755527
při	7159444	podle	5139103	ze	910030	so	742042
bylo	6895481	tam	5098763	mi	882506	ja	735404
ještě	6831030	kde	4922387	ani	849737	roku	734158
až	6802987	jejich	4717935	či	840464	zo	718430
není	6795056	toho	4652283	má	838858	pred	684540
aby	6637007	asi	4644219	však	810090	bola	659994
byla	6411943	ho	4626361	ho	793819	viac	653439

Sample Paragraph With Words Scored

Score = Sum of logarithms of relative frequencies of words in the wordlists

= word score for CS	CS	xCS	SK	xSK	EN	DE	PL	SL	HR	FR
Solo	3.35	3.64	3.34	3.56	4.36	3.92	3.84	4.13	4.23	4.23
Pieces	2.29	2.58	2.33	2.54	4.76	2.75	2.42	2.48	2.72	3.56
for	4.49	4.77	4.47	4.69	7.07	4.67	4.64	4.56	4.93	4.62
La	4.62	4.90	4.55	4.77	4.93	4.80	4.63	4.61	4.77	7.42
Naissance	1.12	1.41	1.05	1.27	1.69	1.32	1.78	1.18	0.82	4.87
de	4.96	5.25	4.92	5.14	5.28	5.15	4.95	4.97	4.93	7.69
L	4.95	5.24	5.01	5.23	4.80	4.74	4.89	4.61	4.74	5.39
Amour	2.59	2.87	2.49	2.71	2.83	2.95	2.52	2.56	2.72	5.30
je	7.16	7.45	7.15	7.37	3.55	5.44	5.77	7.51	7.50	6.76
soundtrackové	1.30	0.00	0.87	0.00	0.00	0.00	0.00	0.00	0.00	0.00
album	4.59	4.87	4.76	4.98	4.74	4.83	4.56	4.78	4.90	4.93
velšského	2.13	0.00	0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00
multi...isty	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Johna	4.07	4.36	3.93	4.15	1.33	1.13	3.92	3.88	4.05	0.56
Calea	1.66	1.95	1.59	1.81	2.13	1.38	1.46	1.54	2.03	1.35
= sum	49.28	49.29	47.21	48.22	47.47	43.08	45.38	46.81	48.34	56.68

Data Processing Steps II

- Tokenisation (Unitok)
- Lemmatisation (Majka)
- Tagging (Majka + Desamb)
- Gender respecting lemma ("veřejné knihovně" → "veřejná knihovna")
- Indexed for search (Manatee/Sketch Engine)
- Precomputed frequencies, word sketches, thesaurus, subcorpora, terms (Manatee/Sketch Engine)

Thanks to my colleague Marek Medveď for running the tagger.

Collecting Wikipedia

- Wiki2Corpus
- November 2017
- Pages & talks

Corpus Vertical Sample

```
<doc id="4235757" src="web17" title="Matěj Kouba - Blog iDNES.cz"
    length="5k-10k" enc_chared="cp1252" crawl_date="2017-11-09 22:53"
    ip="185.17.117.47" url="https://kouba.blog.idnes.cz/?strana=4"
<p heading="no">
<s>
Šel                 k5eAaImAgMnS  jít-v              jít
jsem                k5eAaImIp1nS  být-v              být
tuhle               k6eAd1        tuhle-a            tuhle
za                  k7c7          za-p               za
svou                k3xOyFgFnSc7  svůj-d            svůj
osmdesátiletou      k2eAgFnSc7d1  osmdesátiletý-j    osmdesátiletá
maminkou            k1gFnSc7      maminka-n          maminka
<g/>
.                   kIx.          .-x                .
</s>
</p>
</doc>
```

Corpus Size

- 12,586,415,546 tokens in 35,995,251 documents
- 91 % of tokens from TLD `.cz`

	Total	**Wiki**	**Talk**	**W17**	**W16**	**W15**
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 %

Lexicon Sizes

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

Document Count

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

Token Counts, Type-Token Ratio

Corpus	**Token count**	**Word lexicon**	**TTR**
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

Average Lengths, The Score

Corpus	Avg. tok/doc	Avg. tok/s	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

The-score = the rank of word "the" in a list of lowercased words.

The comparison is based on just a single word and is a bit unfair since word 'the' is a part of the frequency wordlist used to filter csTenTen17.

Keyword Comparison of 2017 Subcorpus To Other Coprora

	csTenTen12		Araneum Bohem.		csSkELL	
R	**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
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
9	trumpa	6.3	mld	6.9	naruto	7.7
10	azure	6.2	hl	6.4	kdyz	7.6
14	sýrii	4.6	atp	4.8	severus	6.7
15	dodavatelský	4.4	ú	4.5	panička	6.5
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

Keyword Comparison of 2017 Subcorpus To SYN 2015

	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

Keyword Comparison of 2016 To 2017 Subcorpora

	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

Word Sketches – “To Grasp A Straw” in csTenTen12

"chytat ... stéblo" of ...		
naděje	26	•••
tráva	13	•••
sláma	3	•••

modifiers of "chytat ... stéblo"		
pověstný	15	•••
příslovečný	3	•••
pomyslný	4	•••

prepositional phrases		
"chytat ... stéblo" v ...	3	•••

Word Sketches – “To Grasp A Straw” in csTenTen17

modifiers of "chytat stéblo"		
pověstný	18	•••
příslovečný	4	•••
pomyslný	4	•••
sebemenší	3	•••

"chytat stéblo" of ...		
naděje	50	•••
tráva	29	•••
záchrana	4	•••
sláma	3	•••
radost	3	•••

subjects of "chytat stéblo"		
tonoucí	66	•••
novinář	3	•••
jídlo	5	•••

prepositional phrases		
"chytat stéblo" v ...	14	•••

Thesaurus – “Antedeluvian” in csSkELL 2.2

	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

Thesaurus – “Antedeluvian” in csTenTen17

	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	středný 4,196

Conclusion & Future Work

Achievement: A new ten-billion-word Czech web corpus

Future Work:

- Correct the tokenisation of abbreviations
- Correct the lemmatisation of foreign words
- Address the part of the corpus from 2016 containing online gambling advertisement spam
- Identify topics and genres of documents

Thank you for your attention!

Photo credit: Kathleen & Ryan Rush