Recent Czech Web Corpora

Vít Suchomel

Natural Language Processing Centre
Faculty of Informatics
Masaryk University

8 December 2012

Introducing czTenTen12

- data gathered in 2011 by crawler SpiderLing
- 20,000 seed URLs $\rightarrow$ 15,000,000 documents gathered
- encoding detection $\rightarrow$ UTF-8
- boilerplate removal by jusText
- tokenized by unitok
- deduplication by onion (7-grams of words, 50 % threshold)
- language detection, extra care of Slovak (6.8 % of docs)
- paragraphs not containing diacritical marks tagged
- dealing with nonwords (extra cleaned version – Marek Grác)
- sentences & tagging by Ajka + Desamb (Pavel Šmerk)
  in 3 days (Pavel Hančar)
  POS, gender, number, case, aspect, modality,. . .
- available at `ske.fi.muni.cz` (free access for MU),
  `sketchengine.co.uk` (free trial access)

Comparison with other Czech corpora

Basic comparison			
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

	number of hits	
corpus	*bavorák*	*email*
SYN2000	24	86
czes2	75	8,549
SYN2010	474	21,897
Hector	1,029	73,781
czTenTen12	1,054	254,548

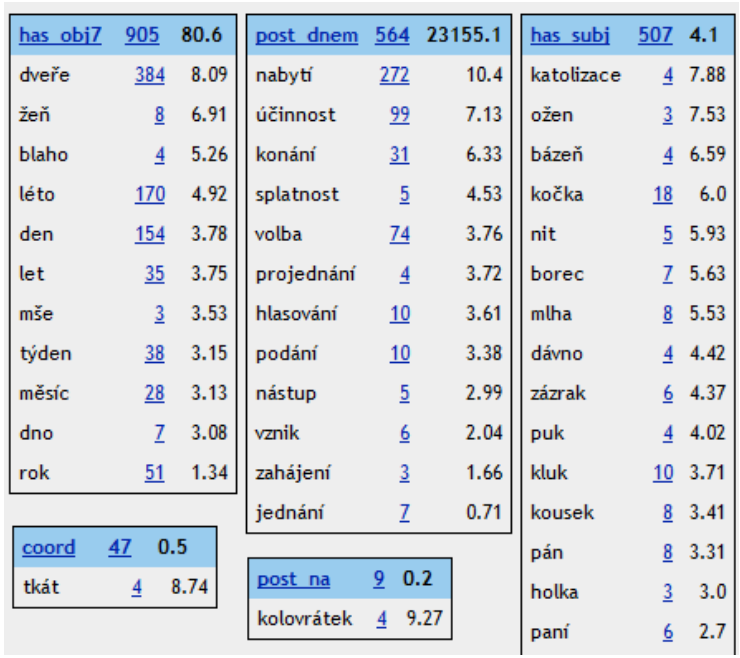
Similarity of corpora

Crosstable of corpora distance				
corpus	czes2	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

key words comparison	most key words
czTenTen12 vs. Hector	*již, lze, oblasti, společnosti, zařízení, této, roce, zde, mohou, rámci, projektu*
Hector vs. czTenTen12	*no, holky, jo, xD, D, blog, teda, taky, já, dneska, sem, jdu, máš*

Word sketches – a bigger corpus yields better collocations

word sketches in SketchEngine



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

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
přist	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

Conclusion

- czTenTen12 – the biggest Czech web corpus for language research
- more Slavonic corpora coming in the future: Polish, Croatian
- interesting topic for further research: studying semantic topics extracted from documents in the corpus