CJDSL001 Korpusová lingvistika (1)

Klára Osolsobě

osolsobe@phil.muni.cz

Experimentální a počítačová lingvistika

O čem budeme mluvit

- Krátký historický exkurz
- Definice korpusu v moderním slova smyslu
- Dva metodologické přístupy k vytěžování korpusu
- Dva pohledy na korpus (lingvista a informatik)
- Filologie a korpusy
- Výuka jazyků a korpusy

Krátký historický exkurz

- Myšlenka korpusu
- Korpusová lingvistika – empirická disciplína
- Data a introspekce
- Technický pokrok

Konkordance – KWIC(**K**ey **W**ord **I**n **C**ontext)

Výskytů: **213** | i.p.m. : **1,76** (vztaženo k celému "omezeni/syn2015") | ARF : **88,32** | Výsledek je promíchán 1 / 6

Výběr řádků: základní | Atributy:

Brno Business	se ho 4379 lidí . Vyplynulo z něj , že	Češi jsou	se svými šéfy poměrně spokojeni . Šéfů , kteří se
Reflex	jsou hlavní autority , pak je to svědectvím , že	Češi jsou	národ-nenárod , jenž se vždycky sjednotí na nějakém kýči ,
Maxim	bezpečnost . Žádný slovenský záchranář si nezapomene rýpnout , že	Češi jsou	blázni . Devět z deseti mrtvých v Tatrách jsou Češi
Týden	, který máme dobýt obchvatem shora , horskými hřebenovkami .	Češi jsou	na Tematíně ostatně dávnými hosty , hrad pokořil už Přemysl
Sport	. Česká vzpoura v Letošní NHL je v polovině a	Češi jsou	vidět ! Po několika hubenějších sezonách je tu konečně vydařený
Reflex	vnímáno hlavně jako neustálý konflikt dvou znepřátelených národů , kde	Češi jsou	vždy ti malí , ale spravedliví , utlačovaní , ale
Mladá fronta Dnes	by to ryzí zázrak . Matematicky to spočítat nelze .	Češi jsou	odkázáni na pomoc shůry . Kromě toho , že sami
Lidé a země	předpisu rakouské armády z roku 1749 . " Proslulí nadšenci	Češi jsou	v Evropě obecně známi jako jedni z nejlepších „ vojáků
Reflex	, že mají zhasínat , pokud jdou z místnosti .	Češi jsou	národem s maloměstským myšlením a jsou závistiví . Pozorujete to
Reflex	koncem dubna byl projednán návrh zákona o její regulaci .	Češi jsou	totiž vůči prostituci značně liberální . A nejen vůči ní
Mladá fronta Dnes	Alsasané ve Francii skoro dodnes . Řekla bych , že	Češi jsou	dnes s vyrovnáváním se s touto částí naší minulosti dál
Týdeník Květy	v hrubém domácím produktu na hlavu je přímo propastný :	Češi jsou	35x bohatší ! Za zmínku určitě také stojí , že
Parlamentní listy	Češi leniví a s tím klesá jejich ochota recyklovat »	Češi jsou	ochotni odnášet vyřazené světelné zdroje maximálně do vzdálenosti 7
Mladá fronta Dnes	tomu není žádný důvod , nic se nemění , ale	Češi jsou	velmi finančně negramotní . Nepřehnalo se to v Česku s
Reflex	naprosto zřejmé , že dává přednost startování auta klikou ?	Češi jsou	někdy trošku záhadní . Taxikář se mě znovu zeptal na
Rodinný dům	několik hollywoodských komedií o Santa Clausovi , testujíce , zda	Češi jsou	již převychováni , takže lze očekávat sérii sebevražd několika program
Respekt	, že oslepl právě po vodce této společnosti . „	Češi jsou	nejen světovými přeborníky v pití piva , ale podle Světové
Metro	na horách , v Paříži nebo překvapivě doma . „	Češi jsou	tedy spokojení nebo bez fantazie , " vyplývá z průzkumu
Parlamentní listy	Dánsko , Švédsko , Německo , Velká Británie) ,	Češi jsou	se 17 % polepšených kuřáků v dolní polovině tabulky .
Lidové noviny	poklesu spotřebitelské poptávky . Není důvodem i to , že	Češi jsou	příliš opatrní a skeptičtí ohledně stavu země a jejího dalšího

PSJČ

---

psjc Příruční slovník jazyka českého

vroucí adj. vařící horký vřelý Vzala čajovou konvičku, hodila tam bezový květ a nalila na to vroucí vody. Vrch. Vroucí vodu napouštěla na odvar čaje ve sklenicích. Mach. Vařil na lihovém kahanu lék. Odbíhaje od vroucího plecháčku k oknu své pracovny, pečlivě pozoroval noční oblohu. John. Jakási jiná, vroucí, tmavá krev byla nalita v žilách těchto lidí. R. Svob. Z nedovřených rtů sálal jí vroucí dech. Čap. Ch. **D** *Zř. vařící se, kolotající, vířící.* Vlny bičované a rozryté tisícerými křižujícími se jizvami a vroucí tisícerými víry. Pašek. **D** *citově opravdový; hluboký, vřelý.* Mluvil tlumeně, vroucím hlasem. Jir. U příležitosti pátého výročí osvobození Československa posílám Vám své vroucí přátelské pozdravy. R. právo. Dívka zvedla kvapně skloněnou hlavu, vroucí pohled utkvěl na okamžik na mladém, statečném muži. Jir. Bůh vroucí modlitby její neoslyšel. Arb. Přitiskla vroucí políbení na jeho ruku. Schulz. Zasloužíš si takto někdy ještě matčin vroucí dík. Zey. Má vroucí duše žije jen tobě, přírodo. Vrch. Ty, od jakživa tak vroucí ctitel všeho venkovského, jsi nevěren svým zásadám? Just. Byl vroucím přívržencem josefinského osvícenství. J. Vlč. Je tu mnoho hluboce procítěného a opravdu vroucího. Vrch.

SEU (**SURVEY OF ENGLISH USAGE**)

BROWN CORPUS
W. Nelson Francis - Henry Kucera

- 1964
- 1. elektronicky zpracovaný korpus
- 1 milion slovních tvarů
- britská a americká angličtina
- pečlivý výběr textů
- vzorky

Definice korpusu v moderním slova smyslu

- Elektronické uložení
- Elektronická přístupnost
- Definovaný obsah (ČEHO) a rozsah (KOLIK)
- Standardní anotace – metada a interpretace jazykových jednotek
- Rychlost, spolehlivost a opakovatelnost vyhledávání a kvantifikace nalezeného

Dva metodologické přístupy k vytěžování korpusu

- Corpus based / korpusem ověřovaný, na korpusu založený výzkum
- Pravidlo/výjimka – otevřený/uzavřený seznam, frekvence
- Corpus driven / korpusem inspirovaný výzkum, korpusem řízený výzkum
- Výzkum kolokací /lexical bundles

Maskulina mají v češtině v gen. pl. koncovku **–ů** (**pánů, hradů, mužů, strojů, předsedů, soudců**). Z tohoto pravidla existují výjimky. Které? Kolik?
[lc!=".*ů" & tag="NN[MI]P2.*"]

za nímž sedí u svého náčiní rybáři . Ostatním vyjma	rozhodčích/rozhodčí/NNMP2-----A-----	je tam po dobu soutěže zakázán vstup . " Letos
ský Beroun - Policisté obvinili osmadvacetiletou chovatelku	koní/kůň/NNMP2-----A-----	z Moravského Berouna na Bruntálsku z týrání zvířat . Podle
v Mosambiku za posledních 50 let postihly zatím asi milion	lidí/člověk/NNMP2-----A-----	. Oficiálně je registrováno kolem 200 obětí , skutečný počet
a Pospěcha spokojen . " Potřebovali bychom však takových	lidí/člověk/NNMP2-----A-----	víc , " poznamenal . Radní chtějí ve Zlíně zrušit
je prý marné . Tištěné ploše dominují názory starostů a	radních/radní/NNMP2-----A-----	, odlišným sdělením je věnováno jen 2,6 % této plochy
, jako 71 . hráče v pořadí . V dresu	Flyers/flyers/NNIP2-----A-----	odehrál 64 zápasů s bilancí 11 gólů a 26 přihrávek
Regio Taxis Bohemia . Ten by zahrnoval asi dvacet tisíc	lidí/člověk/NNMP2-----A-----	, to znamená pětatřicet obcí od Benátek nad Jizerou až
ístě tragédie policejní komisař Emmanuel Adebayo . Z pěti	lidí/člověk/NNMP2-----A-----	, kteří byli nejblíž místu výbuchu , zbyly jen zčernalé
oho ze zkušených ruských letců , podle něhož smrt desítek	lidí/člověk/NNMP2-----A-----	zavinila lidská hamižnost . Pilot Vjačeslav Achremenko potvr
. " Přijeli lidé z tréninkových skupin z Plzně ,	Klatov/klatovy/NNIP2-----A-----	, Přeštic , Rokycan a Třemošné , " uvedl šéf
něj u nás na každých 100 osob připadá téměř 7	nemocných/nemocný/NNMP2-----A-----	a délka nemocenské je mnohdy delší než 30 dní .
Tu pořádá Občanské sdružení Omega ... lame spokojení	lidí/člověk/NNMP2-----A-----	přišlo a stále chodí hodně . Ve středu převezeme první

Celkem: 5056 (102 str.)

	Filtr	word	Frekvence
1.	p/ n	lidí	2 515 537
2.	p/ n	peněz	647 968
3.	p/ n	dní	465 084
4.	p/ n	obyvatel	243 036
5.	p/ n	přátel	125 885
6.	p/ n	koní	102 894
7.	p/ n	hostí	92 521
8.	p/ n	rozhodčích	78 016

	Filtr	word	Frekvence
1.	p/ n	milionů	1 400 417
2.	p/ n	metrů	806 199
3.	p/ n	bodů	725 571
4.	p/ n	hráčů	588 016
5.	p/ n	kilometrů	504 610
6.	p/ n	dolarů	491 535
7.	p/ n	měsíců	475 891
8.	p/ n	miliónů	474 446
9.	p/ n	mužů	453 972

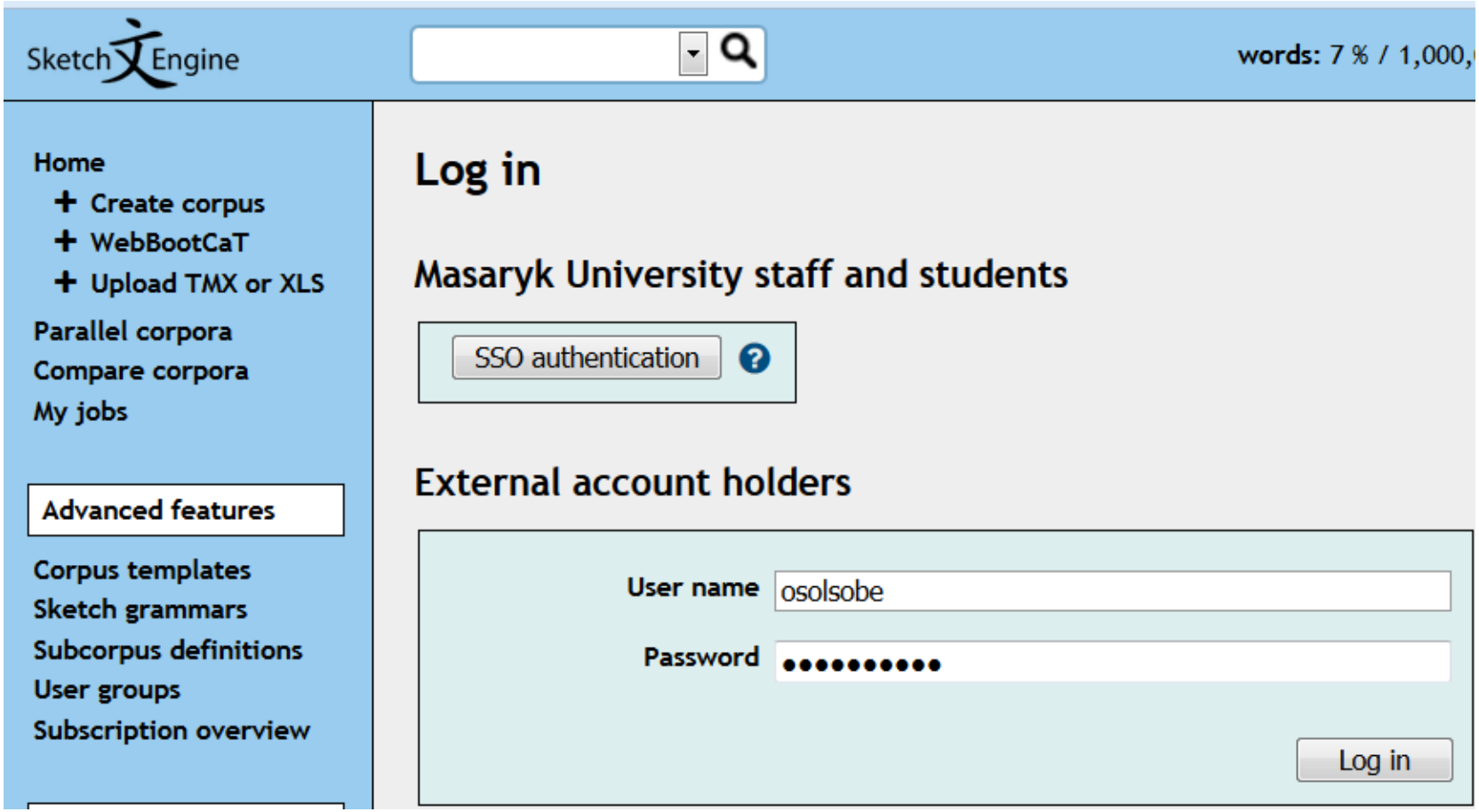
Jaké je mínění o Češích?
Kolokace na pozici 1-3 vpravo od KWIC **<Češi jsou>** seřazené podle míry **MI-score**

	Filtr	lc	Freq ▼	MI	T-score	logDice
1.	p / n	pověstni	3	16.845	1.732	2.843
2.	p / n	dobírkový	4	16.623	2.000	3.258
3.	p / n	nedočkavější	3	14.693	1.732	2.840
4.	p / n	nejateističtějším	5	14.360	2.236	3.572
5.	p / n	pivařský	4	14.108	2.000	3.251
6.	p / n	rovnostáři	5	14.065	2.236	3.570
7.	p / n	rasisti	28	13.830	5.291	5.997
8.	p / n	bytostní	3	13.343	1.732	2.834
9.	p / n	švejci	5	13.326	2.236	3.563
10.	p / n	remcalové	3	13.301	1.732	2.833
11.	p / n	národem	277	13.259	16.642	8.583
12.	p / n	šetřiví	4	13.163	2.000	3.243
13.	p / n	recesisti	3	13.108	1.732	2.832
14.	p / n	náruživými	4	12.899	2.000	3.240
15.	p / n	nejateističtější	4	12.723	2.000	3.237
16.	p / n	smějící	36	12.708	5.999	6.242
17.	p / n	švejkové	3	12.631	1.732	2.827
18.	p / n	přeborníci	23	12.495	4.795	5.642
19.	p / n	studení	9	12.494	2.999	4.372
20.	p / n	slovani	3	12.336	1.732	2.823
21.	p / n	nejsprostší	6	12.329	2.449	3.801
22.	p / n	rasisté	21	12.310	4.582	5.505

Češi jsou národem …

	Filtr	lc	Freq ▼	MI	T-score	logDice
1.	p / n	kutilů	17	17.397	4.123	8.125
2.	p / n	pivařů	19	17.338	4.359	8.094
3.	p / n	houbařů	31	17.186	5.568	8.023
4.	p / n	chatařů	18	16.397	4.243	7.234
5.	p / n	pejskařů	10	15.917	3.162	6.725
6.	p / n	chalupářů	12	15.894	3.464	6.725
7.	p / n	vášnivých	7	15.585	2.646	6.376
8.	p / n	sázkařů	5	15.019	2.236	5.817
9.	p / n	zahrádkářů	5	12.695	2.236	3.607
10.	p / n	milovníků	3	11.531	1.731	2.452
11.	p / n	muzikantů	3	10.551	1.731	1.483
12.	p / n	lyžařů	3	10.513	1.731	1.445

SketchEngine (učo+sekundární heslo)



SketchEngine

- Nástroj disponuje dalšími funkcemi zpracování jednotek (slovních tvarů/lemmat) v korpusech
- Slovní profily (wordsketches) – gramatická kombinovatelnost slov
- Zobrazování slov na základě podobností ve výskytu (thesaurus)

Funkce Word Sketch

- Umožňuje vytvářet vizualizace frekvenčně uspořádaných gramaticky definovaných relací, do kterých vstupuje klíčové slovo v daném korpusu
- Nástroj má zabudována pravidla parciální syntaktické analýzy založené na morfologických značkách
- Tak například na základě toho, že se v bezprostředním levém kontextu substantiva vyskytuje adjektivum, které se shoduje se substantivem v relevantních gramatických kategoriích, je vytvořen seznam a_modifier (adjektivních modifikátorů) typických (s relevantí frekvencí) pro klíčové substantivum)

Word sketch *latina*

latina *(noun)*

Czech Web 2012 (czTenTen12 v9) frekvence = 26,837 (5.29 v milionu)

a_modifier			prec_prep			gen_2			coord			is_obj4_of		
		7.87			32.51			8.79			7.88			6.78
myslivecký ✚	159	7.84	vedle	58	3.23	učebnice ✚	131	7.95	řečtina ✚	746	12.75	sekat ✚	868	11.64
myslivecká latina			vedle latiny ,			učebnice latiny			latiny a řečtiny			sekat latinu		
vulgární	72	7.16	namísto	6	2.69	výslovnost	30	7.86	němčina ✚	228	9.44	tancovat	25	7.91
vulgární latiny			z ✚	2,627	2.60	výslovnost latiny			hebrejština	45	8.96	tancovala latinu		
ciceronský	9	7.11	z latiny			kódování	42	7.55	staroslověnština	33	8.93	učít	24	7.74
středověký ✚	278	6.71	místo	36	2.39	kódování Latin			francouzština ✚	114	8.85	se učil latinu		
středověké latiny			místo latiny			quartier	9	6.96	starořečtina	30	8.83	tančit	31	7.38
akvaristický	14	6.70	kromě	56	2.35	lingua	7	6.58	italština	50	8.64	tančit latinu		
akvaristické latiny			kromě latiny a			znalost ✚	236	6.43	čeština ✚	113	8.34	učit ✚	113	6.20
scholastický	8	6.33	včetně	43	1.56	znalost latiny			funk	26	8.02	učit latinu		
námořnický	15	6.33	včetně latiny a			výuka ✚	148	6.35	funku a latiny			vytlačovat	9	6.11
zkomolený	8	6.29	do ✚	994	1.23	výuky latiny			aramejština	17	7.92	vyučovat	25	5.97
hovorový	22	6.19	do latiny			profesor ✚	117	6.26	aramejštině a latině			ovládat	98	5.67
hovorová latina			oproti	9	1.21	profesor latiny			španělština	34	7.68	ovládal latinu		
humanistický	18	5.64	prostřednictvím	13	1.09	profesorka	16	6.16	ruština	35	7.58	vytlačit	11	5.60
archaický	10	5.23	prostřednictvím latiny			profesorka latiny			jazz	34	7.54	vytlačila latinu		
rybářský	61	5.21	v ✚	3,495	1.08	znalec	30	6.06	jazzu a latiny			studovat ✚	105	5.42

Zrada v podobě homonymie (paronymie)
jazz/tancovat/tačit

Funkce Thesaurus (zobrazení podobných slov)

- Na základě porovnání kontextů je vytvořen seznam a vizualizace slov, která mají podobné (gramaticko-lexikální) kontexty

latina *(noun)*
Czech Web 2012 (czTenTen12 v9) frekvence = 26,837 (5.29 v milionu)

Lemma	Skóre	Frekvence
francouzština	0.367	31,593
ruština	0.349	25,527
němčina	0.336	62,767
řečtina	0.320	10,714
španělština	0.316	19,214
arabština	0.281	8,834
italština	0.277	11,852
hebrejština	0.270	7,170
čínština	0.266	8,672
polština	0.262	10,041
slovenština	0.247	14,334
japonština	0.236	7,430
angličtina	0.228	212,669
čeština	0.192	227,991
romština	0.188	4,026
matematika	0.183	88,091
maďarština	0.183	4,228
dějepis	0.183	26,800
portugalština	0.160	3,558
zeměpis	0.156	14,562
švédština	0.151	2,735

Sketch rozdíl (vizualizace kontextu dvojice): ***čeština/latina***

- Společné kontexty (bíle)
- Kontexty typické pro každý
- člen dvojice (zeleně
- a červeně podbarvené)

a_modifier	2,113	29,737	0.08	0.13
myslivecký	159	0	7.8	--
ciceronský	9	0	7.1	--
akvaristický	14	0	6.7	--
scholastický	8	0	6.3	--
námořnický	15	0	6.3	--
rybářský	61	0	5.2	--
vulgární	72	25	7.2	4.2
středověký	278	58	6.7	4.1
znalý	22	23	5.1	4.0
zkomolený	8	42	6.3	5.5
vytříbený	12	42	5.2	5.2
humanistický	18	101	5.6	6.4
archaický	10	100	5.2	6.5
hovorový	22	854	6.2	9.5
mluvený	23	581	4.8	8.4
srozumitelný	8	269	2.9	7.1
spisovný	23	5,797	4.7	11.7
demo	0	112	--	6.5
bezchybný	0	149	--	6.6
jadrný	0	125	--	6.9
psaný	0	359	--	7.1
přeložený	0	230	--	7.5
obecný	0	1,973	--	7.7
znakovaný	0	462	--	9.0
lámaný	0	554	--	9.1

Dva pohledy na korpus (lingvista a informatik)

- Nástroje NLP a korpusy
- Konverzní programy, vertikál, tokenizér
- Korpusové manažery
- Automatické analyzátory
- Lingvistické interpretace v korpusech

Vyhledání slovního tvaru *jít*

" Můj partner na rakovinu zemřel , protože se bál	jít	k lékaři . Z vlastní zkušenosti tedy vím , jak
, “ uvedl Malášek . Nevyloučil možnost , že může	jít	o vojáka nebo policistu . K druhé loupeži došlo jen
pádu . ^ * S výjimkou případů , kdy má	jít	k veterináři a vy se ji marně snažíte napěchovat do
mrazu . (Lidovci už ztrácejí i Moravu . Zkusí	jít	do měst a zezelenat Jihomoravští lidovci získali v komunálních volbách
patriotům se změna názvu zřejmě líbit nebude , nemusí však	jít	o řešení trvalé . Vše nyní i v budoucnosti závisí
. Tyto potíže se léčí laserem . Po operaci může	jít	pacient hned domů Prof. MUDr. Pavel Kuchynka , CSc. ,
v žádném případě . Je pravda , že jsme mohli	jít	na Švédsko a Kanadu . A nebudu lhát , že
v jedenáctimilionové zemi . ATÉNY „ Nevíme , kam máme	jít	, přijeli jsme lodí z Lesbosu , moje žena je
více Dominiků , než bývá obvyklé . A nemuselo hned	jít	o zapálené hokejové fanoušky . Stačí , že se jméno
unie přijala směrnici o nedovolené podpoře podnikání , která neumožňuje	jít	s nabídkou pod určitou minimální částku , " vysvětlil mluvčí
řadových odborářů , s jakým argumentačním mandátem k němu mají	jít	. Doslova v odborářském šoku jsem po té , co
trasy k nejzajímavějším místům parku . Můžete použít lodě ,	jít	pěšky nebo jet džípem a bahnem se dobrodit až k

Vyhledání lemmatu *jít (KWIC+lemma+tag)*

v Ledči . Pokud dva tři dny pořádně prší ,	jde/jít/VB-S---3P-AA---I	začít pár kilometrů výše proti proudu a zdolat Stvořidla .
to přece dělal baron Prášil ! Americký kaskadér David Smith	jde/jít/VB-S---3P-AA---I	v jeho stopách - o víkendu tenhle bláznivý kousek předvedl
takové se mezi nimi možná občas najdou - , ale	jde/jít/VB-S---3P-AA---I	o spory , jež jdou nutně ruku v ruce s
ale také ceny za téměř třicet tisíc korun . "	Jde/jít/VB-S---3P-AA---I	opravdu o nálož cen kvalitních značek , " řekl pořadatel
připlácet , když chce státní úřad . Zvlášť , když	jde/jít/VB-S---3P-AA---I	o peníze , které tak jako tak od státu předtím
v baru , bowling . Něco , co lidi přiměje	jít/jít/Vf--------A----I	tam , a ne trávit čas jinak . Vesecko nabízí
s onou šaškárnou . Upevňuje se klamné přesvědčení , že	jde/jít/VB-S---3P-AA---I	o replikaci bytostí . Že je to něco jako nesmrtelnost
ání firemního mobilního telefonu a automobilu k soukromým účelům	jdou/jít/VB-P---3P-AA---I	ruku v ruce s rostoucími ekonomickými výsledky společnosti . Zkrátka
filmy promítají . „ Máme rozpis , podle kterého měla	jít/jít/Vf--------A----I	kopie filmu od nás do Sezimova Ústí . Bohužel jsem
byl opilý . Tehdy mi řekl , že když to	nepůjde/jít/VB-S---3F-NA---I	po dobrém , půjde to po zlém , “ vypověděla
na Liberec ? No , dlouho to vypadalo , že	jdeme/jít/VB-P---1P-AA---I	na Karlovy Vary , protože Litvínov v posledním kole vyrovnal
zjevně mladší a nezdravě agresívní . " Jsem ženská a	jdu/jít/VB-S---1P-AA---I	z práce . " odpověděla jsem poněkud mimo . "
v koupelně a přistihla je tam Kate , když si	šla/jít/VpFS---3R-AA---I	upravit nalíčení . Pak ale Pippa začala chodit s bankéřem
neexistující exoty a kdo všechno o tom podvůdku věděl a	nešel/jít/VpMS---3R-NA---I	to oznámit na policii . Ztratit paměť je , milí
, protože se jedná o klasické inflační peníze , které	jdou/jít/VB-P---3P-AA---I	v podstatě do stavebních firem . Mnohem lepší podle mne
= hrůza ! Jaké máte zkušenosti s českými silnicemi ?	Nejde/jít/VB-S---3P-NA---I	mi dost dobře do hlavy stav českých dálnic , které
v síti STEP ( ) . HLEdEJtE spoLEčnou cEstu Pokud	jde/jít/VB-S---3P-AA---I	o předkládání projektů na odstraňování starých ekologických zátěží , nejbližší
. Celý dospělý život jsem měla největší přání - aby	šel/jít/VpMS---3R-AA---I	bolševik od válu . A to se mi splnilo .

desambiguace

- Pánové, *nežeňte se*
- Nemluv a *rožni*.
- Jan *je osel*.

Víceznačné tvary

- *nežeňte/(ne)hnat/V*
- *nežeňte/(ne)ženit/V*
- *se/se/P*
- *se/s/R*
- *rožni/rožnit/V*
- *rožni/rozžehnout/V*
- *rožni/rožeň/N*
- *je/být/V*
- *je/on/P*
- *osel/osel/N*
- *osel/osít/V*

hnát/ženit

pracujte za byt a stravu . Ale nevdávejte se a	nežeňte/hnát/Vi-P---2--N----I se/se/P7--4----------	jen kvůli tomu , že chcete uniknout z dor
sexu Shrnutí těchto vědeckých poznatků : Nejezděte na kole a	nežeňte/hnát/Vi-P---2--N----I se/se/P7--4----------	s učitelkou . Potkáte - li ženu , která krá
legendární Knoflenkou . Takže se mějte a smějte (a	nežeňte/hnát/Vi-P---2--N----I se/se/P7--4----------	a nevdávejte) . KDO S KÝM KDE
vážný důvod , proč tam chce jet . Tvůrci :	Nežeňte/hnát/Vi-P---2--N----I se/se/P7--4----------	do ženění HEJNICE Pot
takovýchto " žertíků " neustále a iv jiném prostředí .	NEŽEŇTE/hnát/Vi-P---2--N----I SE/se/P7--4----------	ZA SLEVOU , ŽÁDNÁ NENÍ Jedn
Uvědomte si , že vaše zásoba energie je omezená .	Nežeňte/hnát/Vi-P---2--N----I se/se/P7--4----------	z jedné činnosti do druhé . 7 . V
Udo Pollmer říká : " Vyhoďte z bytu televizi a	nežeňte/hnát/Vi-P---2--N----I se/se/P7--4----------	. Svatba je počátkem tukového obalovár
ventilaci - či pootevřete e dveře . Když skončíte ,	nežeňte/hnát/Vi-P---2--N----I se/se/P7--4----------	sprchovat - studenou vodou , ale zůstaň
to , co vyděláte hned dál investujte . 3 .	Nežeňte/hnát/Vi-P---2--N----I se/se/P7--4----------	do nejistých investic Nezkoušín
požehnaného Čechova , jenž pravil , cituji : „ '	Nežeňte/hnát/Vi-P---2--N----I se/se/P7--4----------	, pokud se bojíte osamělosti . ' " <p:
se mohli vrátit k zaříkadlům , která přivolávají lásku .	Nežeňte/hnát/Vi-P---2--N----I se/se/P7--4----------	však hned do vážných známostí , nejdřív
na cizí . Sebestian Roch Chamfort Bojíteli se samoty ,	nežeňte/hnát/Vi-P---2--N----I se/se/P7--4----------	! Anton Pavlovič Čechov Bůh stv
Uvědomte si , že vaše zásoba energie je omezená .	Nežeňte/hnát/Vi-P---2--N----I se/se/P7--4----------	z jedné činnosti do druhé . 7 . V
světě vidět a chovat vlastní dítě . Nepodléhejte trendům a	nežeňte/hnát/Vi-P---2--N----I se/se/P7--4----------	na sál jen proto , že kamarádi byli . Zept
vědění je chromá . Věda bez náboženství je slepá .	Nežeňte/hnát/Vi-P---2--N----I se/se/P7--4----------	za detaily . Nechť je vám v myšlenkách i
růst zajistí . 5 . Buďte trpěliví a nevzdávejte to	Nežeňte/hnát/Vi-P---2--N----I se/se/P7--4----------	do velkých zakázek bezhlavě . Získávejt
20.00 hodin kino v Hejnicích . Motto filmu zní :	Nežeňte/hnát/Vi-P---2--N----I se/se/P7--4----------	do ženění . Až na svatební cestě Eddie z

být/on

Není to nic moc , ale Millicent to nepozná .	Je/být/VB-S---3P-AA---I osel/osel/NNMS1-----A-----	. Znova - que voulez-vous ? " Prozpěvovaly kvulevu ,
Titulní strana Šípu , která o premiérovi hlásala , že	je/být/VB-S---3P-AA---I osel/osel/NNMS1-----A-----	, či ta , kde trenér Brückner drží před začátkem
a profous považuje za svou povinnost říci mu , že	je/být/VB-S---3P-AA---I osel/osel/NNMS1-----A-----	. Mužstvo jde na těžké cvičení do hor , ale
Nebudu dělat to , co řekne Manuel . „ Manuel	je/být/VB-S---3P-AA---I osel/osel/NNMS1-----A-----	, Pere ! Osel je zvíře a ty přece nebudeš
mějský jarmark . V roce 1616 následovala hra Ďábel	je/být/VB-S---3P-AA---I osel/osel/NNMS1-----A-----	, v níž ani ďábel nestačí na zlo lidí ,
ocourskovský sedlák osít dvě míry pole krupicí . Kdo	je/být/VB-S---3P-AA---I osel/osel/NNMS1-----A-----	, nechť se přihlásí . V říčce , do které
o rozesmálo všechny . Mě tedy nejvíc překvapilo , že	je/být/VB-S---3P-AA---I osel/osel/NNMS1-----A-----	chytřejší než kůň . Po dotazech , které jsme měli
dy smím použít přejemnělého výrazu - jaký ten druhý	je/být/VB-S---3P-AA---I osel/osel/NNMS1-----A-----	. Ale ne tak Evženka . Ta , když vypravuje
kde se před drátěným plotem něco pohybovalo . " To	je/být/VB-S---3P-AA---I osel/osel/NNMS1-----A-----	! " zajásala Julie a už se k němu rozeběhla
: A co dále , Baltazare , kde hlavním hrdinou	je/být/VB-S---3P-AA---I osel/osel/NNMS1-----A-----	nebo u S. Paradžanova : Barva granátového jablka , kde
osla a pro jistotu pod obraz napsal : " Toto	je/být/VB-S---3P-AA---I osel/osel/NNMS1-----A-----	. " Bedlivější analýzou vyprávěných anekdot lze zjistit , že
to na tebe narafičil , ty osle ! Román Gaius	je/oni/PPFP4--3-------- osel/osel/NNMS1-----A-----	ukazuje , že Henry Winterfeld psal už před půlstoletím lepší
Trutnovska z obce Mladé Buky , že v ulici Tmavá	je/být/VB-S---3P-AA---I osel/osel/NNMS1-----A-----	. „ Policisté díky místní znalosti vypátrali jeho majitele a

rožeň/rožnit/rozžehnout

NY . , O. LI . V . Y ,	ROŽNI/rožeň/NNIP7-----A-----	, K. I . Š , Pa . Le .
to vím jistě . Z nářečí znám třeba slovo „	rožni/rožeň/NNIP7-----A-----	“ . Ester Hozáková , 6 let , FrýdekMístek :
k paní Haleové , „ že s těmi hodinami ,	rožni/rožeň/NNIP7-----A-----	a planetáriem máte doma malé muzeum . “ „ Jestli
jste naučila vy jeho ? Různé moravismy jako šufánek nebo	rožni/rožeň/NNIP7-----A-----	. Je nevěra tak zajímavá , že o ní teď
nás chodí dívat místní . Je tam obrovský krb s	rožni/rožeň/NNIP7-----A-----	, ve kterém se peče drůbež - od bažanta přes
je jen krátký úryvek z nabídky mas na grilu i	rožni/rožeň/NNIP7-----A-----	, které Farma nabízí . Mezi omáčkami , přílohami a
pro služebnictvo sestoupil do kuchyní , kde skomíral oheň pod	rožni/rožeň/NNIP7-----A-----	bez masa . Na zemi vedle stolu na porcování masa
pražsky , což rozhodně popírám ! Děti říkají fajne ,	rožni/rožeň/NNIP7-----A-----	místo rozsviť , štrampliky místo punčocháče . Opakují slova ,
, ale také teplá kuchyně s grily , rošty a	rožni/rožeň/NNIP7-----A-----	, kde gastronomické poklady před našima očima do zlatova dozrávaly
místnosti Dechem řádky staletí s krápníky sazí z ohňů pod	rožni/rožeň/NNIP7-----A-----	prasat či skopců hovoří centrální černá kuchyně , kam by
sklad plný všelikého harampádí a jen mohutné topeniště krbu s	rožni/rožeň/NNIP7-----A-----	a naproti stojící sporák byly důkazem , že stojím v
to pyšný . V Praze pořád používám slovo žufánek nebo	rožni/rožeň/NNIP7-----A-----	. A nerozumějí mi , “ prozradil . Když Okamura
, ale také teplá kuchyně s grily , rošty a	rožni/rožeň/NNIP7-----A-----	, kde gastronomické poklady před našima očima dozlatova dozrávaly .

Filologie a korpusy

- Obecné a specializované korpusy
- Příklady z českého prostředí
- Tvorba vlastního korpusu

ÚČNK http://ucnk.ff.cuni.cz/cs/

- akademický projekt 1994
- systematicky mapuje češtinu i další jazyky
- po bezplatné registraci otevřeny všem zájemcům

Korpusy ÚČNK

https://kontext.korpus.cz/first_form?corpname=omezeni%2Fsyn2015

Specializovaný - příklad

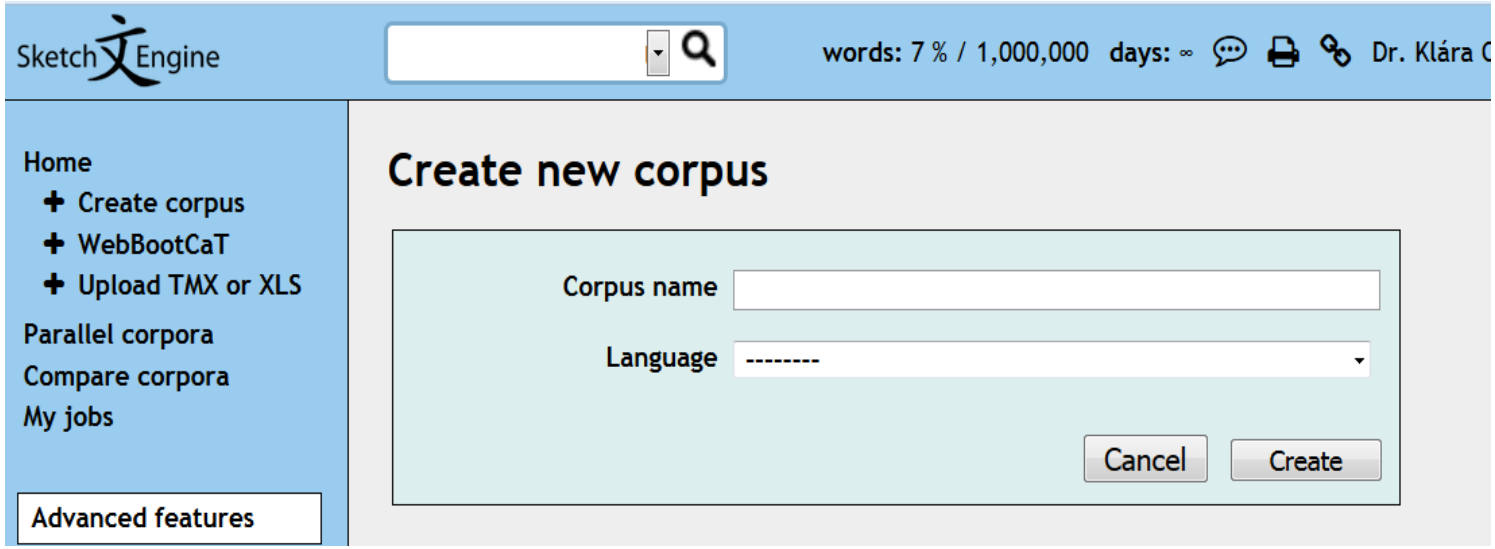
- Korpus Jerome je **jednojazyčný srovnatelný korpus** (*monolingual comparable corpus*) speciálně sestavený pro zkoumání překladové češtiny (tedy textů přeložených do češtiny z jiných jazyků) v porovnání s češtinou nepřekladovou (původní česky psanou).

autorský

- Karel Čapek
- Bohumil Hrabal
- Korespondence Karla Havlíčka Borovského

Tvorba vlastního korpusu - DIY

- https://nlp.fi.muni.cz/cs/JakVytvoritKorpus1
- https://ske.fi.muni.cz/auth/create_corpus/



Výuka jazyků a korpusy

- Metoda DDL (Tim Johnes)
- Žákovské korpusy (Learner Corpora)
- Učebnicové korpusy

Žákovské korpusy

- Texty mluvčích L2
- Výzkum interlanguage
- Zpětná vazba na základě analýzy chyb
- Sledování růstu jazykových kompetencí

Učebnicový korpus Učko

1	Adamovičová, A. – Ivanovová, D. (2007): *Basic Czech I, II.* Praha: Karolinum.
2	Bischofová, J. et al. (2007): *Čeština pro cizince a azylanty B1.* Brno: SOZE.
3	Bořilová, P. – Holá, L. (2011): *Česky krok za krokem 2.* Praha: Akropolis.
4	Cvejnová, J. (2011): *Česky, prosím.* Praha.
5	Čechová, E. – Remediosová, H. (2005): *Chcete mluvit česky? 1, 2* Liberec: Harry Putz.
6	Froulíková, L. (2008): *Adam a Eva v Českém ráji*. Praha: Academia.
7	Hádková, M. (2005): *Čeština pro cizince a azylanty A1, A2.* Brno: SOZE.
8	Holá, L. (2006): *New Czech Step by Step*. Praha: Akropolis.
9	Holá, L. (2010): *Čeština Express 1, 2.* Praha: Akropolis.
10	Kestřánková, M. et al. (2010): *Čeština pro cizince B1.* Brno: Cpress.
11	Matula, O. (2007): *Český den. Kurz českého jazyka pro azylanty navazující na Manuál pro učitele českého jazyka pro cizince bez znalosti latinky.* Praha: Člověk v tísni o.p.s., Projekt Varianty.
12	Nekovářová, A. (2006): *Čeština pro život - 15 moderních konverzačních témat.* Praha: Akropolis.
13	Parolková, O. (2004): *Czech for foreigner s 1, 2.* Praha: Bohemika.
14	Pintarová, M. – Režková, I. (1995): *Communicative Czech. Elementary Czech. Intermediate Czech*. Praha: Univerzita Karlova.
15	Rigerová, K. (2000): *Czech for Everyone*. Praha: K. Rigerová.
16	Štindl, O. (2008): *Easy Czech. Elementary.* Praha: Akronym.
17	Štindlová, B. (2008): *Česky v Česku 1, 2*. Praha: Ústav jazykové a odborné přípravy UK: Akropolis.

chodit do/na (porovnání obecného a učebnicového korpusu)

Celkem: 17391 (348 str.)

	Filtr	word	lc [lowercase word]	Frekvence
1.	p/ n	do	školy	21 199
2.	p/ n	do	práce	15 196
3.	p/ n	na	procházky	4 640
4.	p/ n	do	kina	3 661
5.	p/ n	do	kostela	2 940
6.	p/ n	do	posilovny	2 785
7.	p/ n	do	školky	2 693
8.	p/ n	na	fotbal	2 153
9.	p/ n	do	divadla	2 104
10.	p/ n	do	lesa	2 063
11.	p/ n	do	zaměstnání	1 808
12.	p/ n	na	tréninky	1 629
13.	p/ n	do	hospody	1 540
14.	p/ n	na	pivo	1 534
15.	p/ n	na	houby	1 380

		lemma	word	Frekvence
P	N	chodit na	procházky	5
P	N	chodit do	kina	4
P	N	chodit do	školy	3
P	N	chodit do	divadla	3
P	N	chodit na	oběd	2
P	N	chodit na	dlouhé	2
P	N	chodit na	balet	2
P	N	chodit do	kostela	2
P	N	chodit na	různé	1
P	N	chodit na	ryby	1
P	N	chodit na	lekce	1
P	N	chodit na	koncerty	1
P	N	chodit na	houby	1
P	N	chodit na	hodinu	1
P	N	chodit na	dlouhou	1
P	N	chodit na	brigádu	1
P	N	chodit do	přírody	1
P	N	chodit do	práce	1
P	N	chodit do	lesa	1
P	N	chodit do	kin	1
P	N	chodit do	hospody	1

Závěr

- Historie a současnost – technický pokrok a metodologické přístupy
- Rychlost- spolehlivost – opakovatelnost experimentu
- Zdroje nespolehlivosti
- Co je k dispozici a co si mohu sám udělat
- Na co nezbyl čas

Děkuji vám za pozornost