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.

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

https://www.korpus.cz/kontext/view?maincorp=syn2015&viewmode=kwic&pagesize=40&attrs=word%2Clemma_lc%2Ctag&attr_vmode=visible-kwic&base_viewattr=word&structs=doc&refs=%3Ddoc.title&q=~twWA0sqO4muC

Výskytů: **645** | i.p.m.: **5,34** (vztaženo k celému korpusu) | ARF: **250,56** | Výsledek je promíchán 1

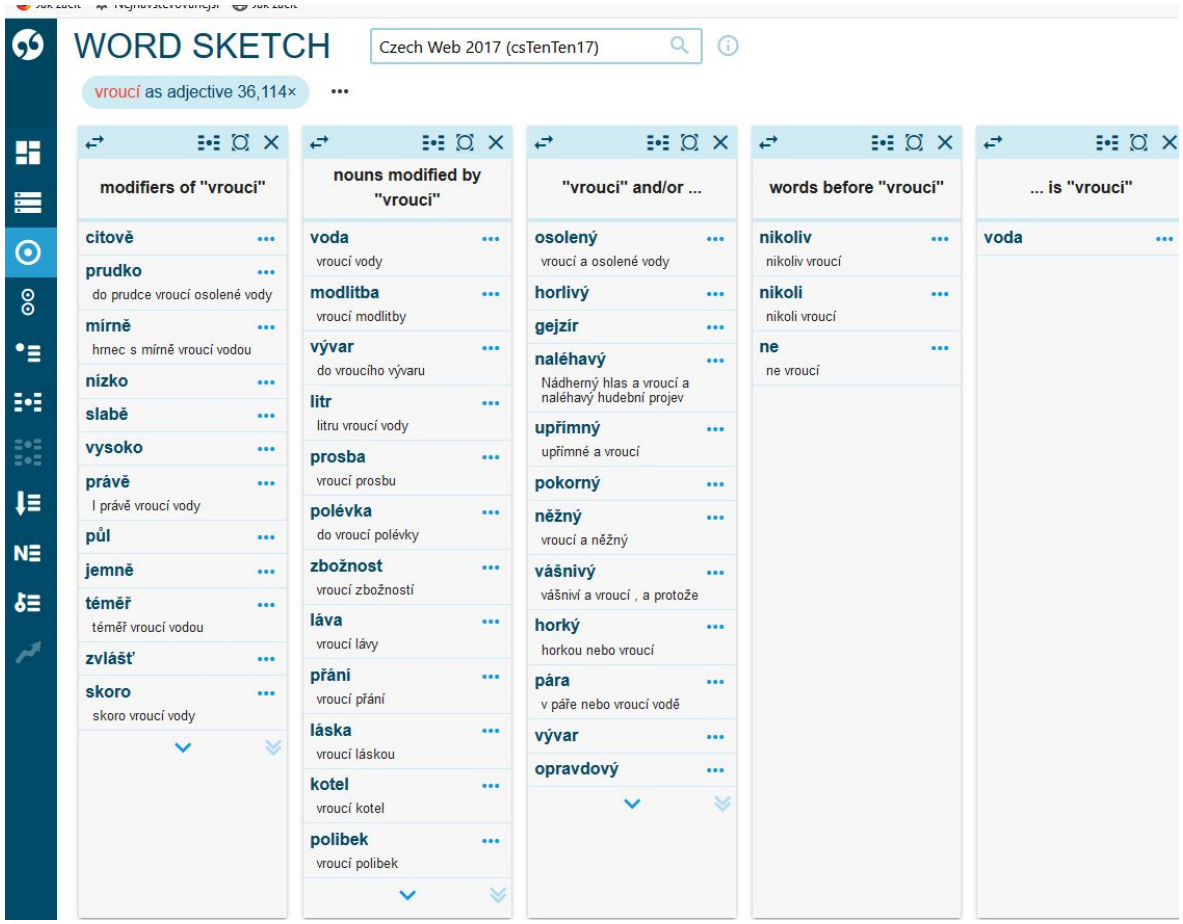
Výběr řádků: základní

F.O.O.D.	. Tvořte knedlíčky , vařte je asi 5 minut ve	vroucí/vroucí/AAFS6----1A-----	osolené vodě a pak je vložte do polévky spolu s
Deníky Moravia	, ani ruské auto už nejde nastartovat . - 39°C	Vroucí/vroucí/AAFS1----1A-----	atmosféra v Kongresu zamrzne . Rusové si zapínají vrchní knoflíky
Mladá fronta Dnes	Nebo se vrhnu po televizním ovladači . A při šumění	vroucí/vroucí/AAFS2----1A-----	vody v trubkách topení se nechám konejšit pouhým vědomím ,
S elegancí ježka	kdy jsme byli zbaveni jha nemoci . Cítila jsem jeho	vroucí/vroucí/AAFS4----1A-----	ruku ve své a vnímala vibrace rozkoše , které v
Fajn život	kůry s 1 lžící květů heřmánku , zalijte 500 ml	vroucí/vroucí/AAFS2----1A-----	vody a nechte 30 minut vyluhovat . Sceďte a nalijte
Pátý elefant	by se mohlo stát člověku , který by padl do	vroucího/vroucí/AAIS2----1A-----	kotle pod vodopádem s ostrým kusem železa připevněným k tělu
Příhody z bezprostředního neskutečna; Zjizvená srd...	životem jsem uvnitř žil i jiný , intimní život -	vroucí/vroucí/AGIS1-----A-----	, milovaný a utajovaný jako nějaké úžasné a fantastické niterné
In magazín	Poté přidáme ještě tři lžíce oleje a zamícháme . Do	vroucí/vroucí/AAFS2----1A-----	osolené vody vložíme připravené gnocchi , a než se udělají
Neobyčejný benediktin	když přijeli do spícího města , vyrušil Cadfael převora z	vroucích/vroucí/AAFP2----1A-----	modliteb zvláštní otázkou . " Otče , měl někdo z
Swamplandie	sirka , kterou škrtneš sám o sebe – maličké ,	vroucí/vroucí/AANS4----1A-----	volání – tak jsem pak asi mohla říct cokoliv ,
Aha! neděle	i zastudena . Do šálku dejte hrst kopřiv a zalijte	vroucí/vroucí/AAFS7----1A-----	vodou . Nechte několik minut louhovat . Pak sceďte a
Rytmus života	vyjmeme je a necháme okapat . 3 Nudle uvaříme ve	vroucí/vroucí/AAFS6----1A-----	osolené vodě , pak je přes síto propláchneme studenou vodou
Strážkyně krve	. Odhodlaně jsem se zadíval z okna , než Mab	vroucí/vroucí/AAFS7----1A-----	vodou zalila čaj v konvici . Odnesla ji spolu se
Zloděj kufrů	se objevuje zapomenutá německá velikost , všem na očích ve	vroucích/vroucí/AANP6----1A-----	německých srdcích . Zásluha NSDAP o německou vědu , německé
Hráčský instinkt	pomalu a opatrně do vany . Horká voda pálila jako	vroucí/vroucí/AAIS4----1A-----	olej , nesnesitelnější než všechno , co se dosud stalo

Seznam lemmat v kolokaci <1,1> vpravo od KWIC

	Filtr	lemma_lc	Freq	MI ▼	T-score	logDice
1.	p / n	osolený	67	15.258	8.185	11.152
2.	p / n	spařit	3	12.499	1.732	7.050
3.	p / n	ml	94	12.250	9.693	9.499
4.	p / n	přelít	7	11.043	2.644	7.501
5.	p / n	zalít	17	10.583	4.120	7.678
6.	p / n	šálek	20	10.506	4.469	7.669
7.	p / n	vývar	8	10.190	2.826	7.088
8.	p / n	modlitba	19	10.178	4.355	7.388
9.	p / n	láva	3	10.058	1.730	6.390
10.	p / n	voda	315	9.867	17.729	7.338
11.	p / n	oddanost	3	9.821	1.730	6.279
12.	p / n	litr	15	9.381	3.867	6.661
13.	p / n	polévka	9	8.933	2.994	6.172
14.	p / n	kotel	7	8.890	2.640	6.073
15.	p / n	olej	14	8.337	3.730	5.712
16.	p / n	nalít	5	8.260	2.229	5.469
17.	p / n	tekutina	5	8.046	2.228	5.291
18.	p / n	vodní	13	7.969	3.591	5.362
19.	p / n	polibek	3	7.936	1.725	5.064
20.	p / n	přátelství	3	7.605	1.723	4.800
21.	p / n	čaj	7	7.601	2.632	4.955
22.	p / n	přání	6	7.357	2.435	4.713
23.	p / n	roztok	3	7.183	1.720	4.446
24.	p / n	káva	6	7.164	2.432	4.535
25.	p / n	mléko	4	6.704	1.981	4.065

Slovní profil



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 potv
. " 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 ... lemo 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?

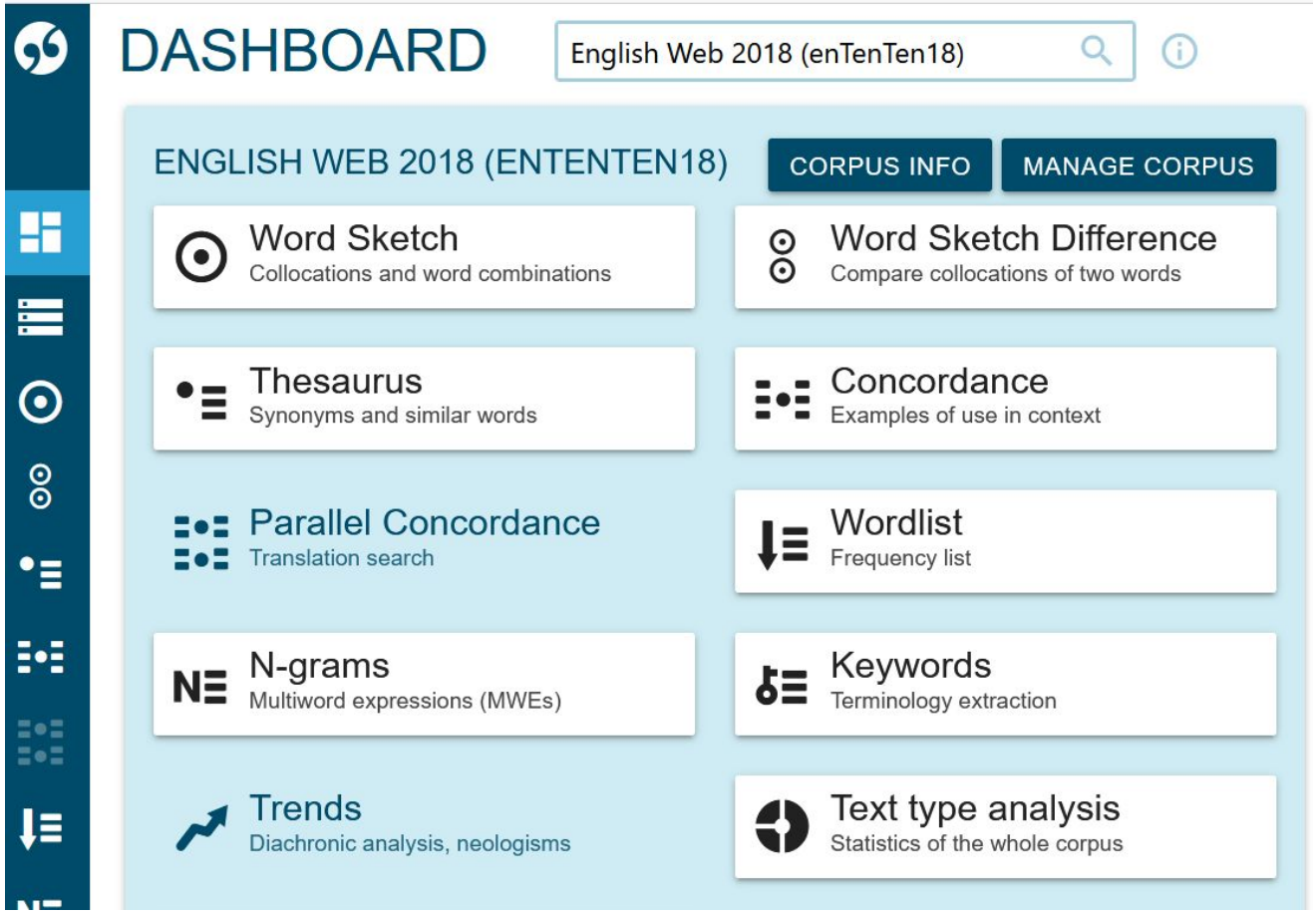
Kolokace na pozici 1-3 vpravo od KWIC **<Češi jsou>** seřazené podle míry **MI-score**

	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)



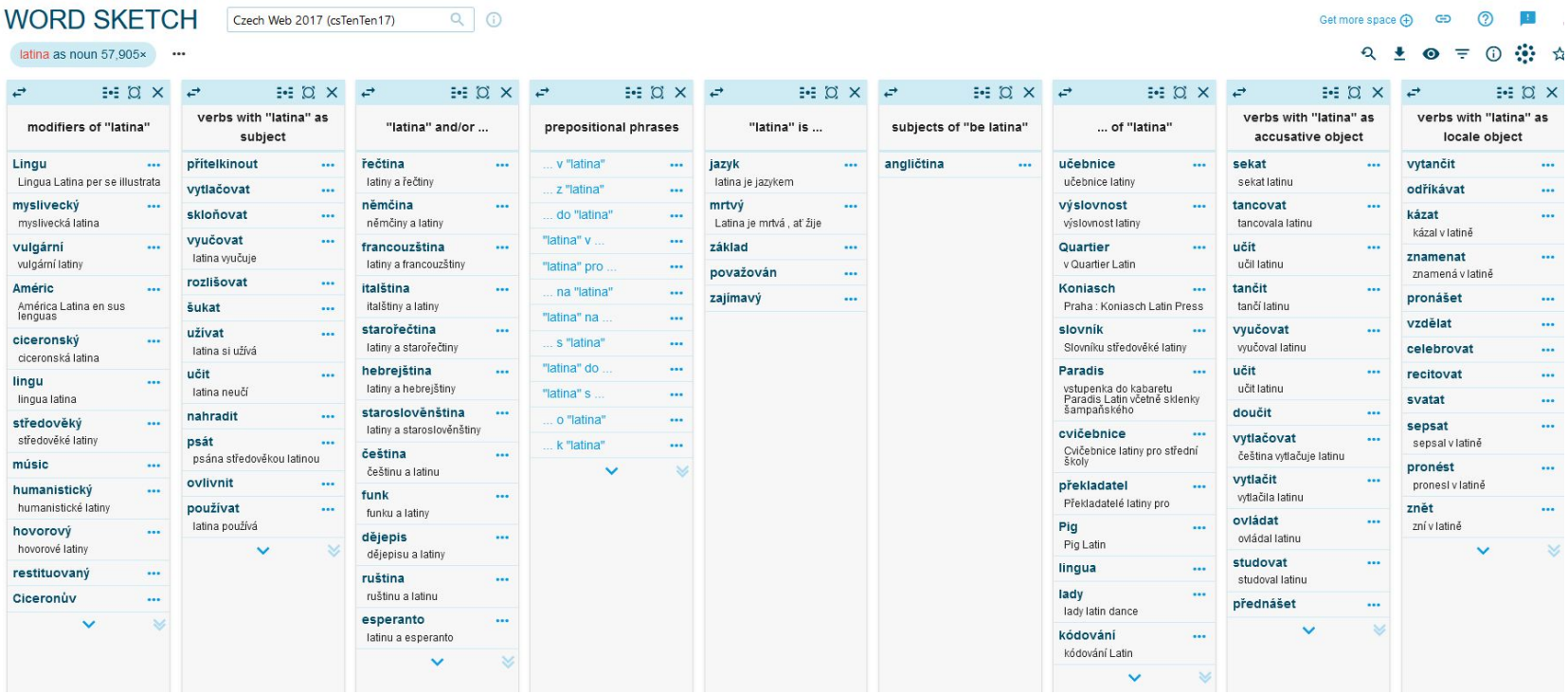
SketchEngine

- 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)

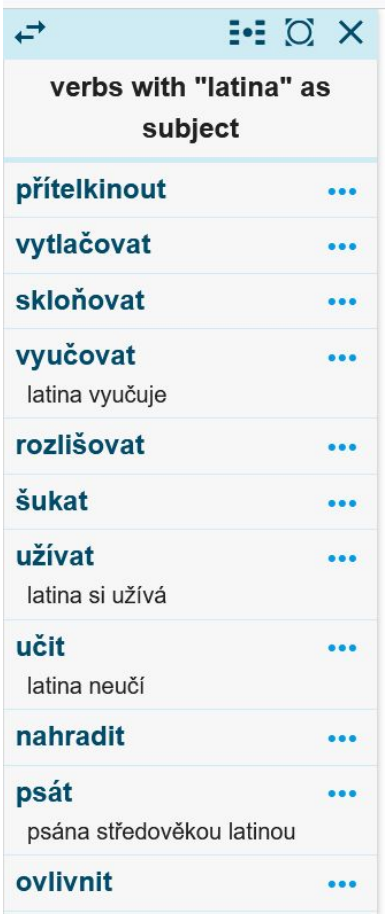
Funkce Word Sketch

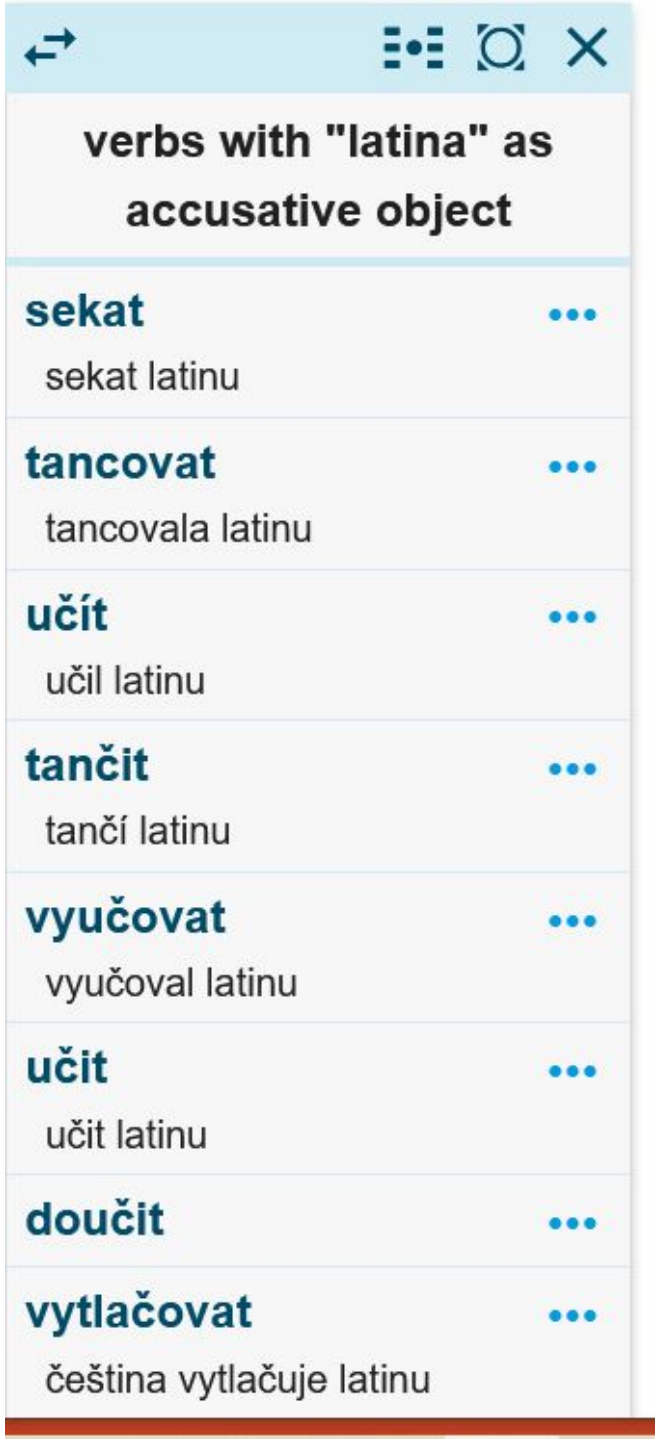
- Umožňuje vytvářet vizualizace frekvenčně uspořádaných gramaticky definovaných relací, do kterých vstupuje klíčové slovo v daném korpusu
- Nástroj má zabudována pravidla parciální syntaktické analýzy založené na morfologických značkách
- Tak například na základě toho, že se v bezprostředním levém kontextu 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*



Co překvapí a proč se objeví?





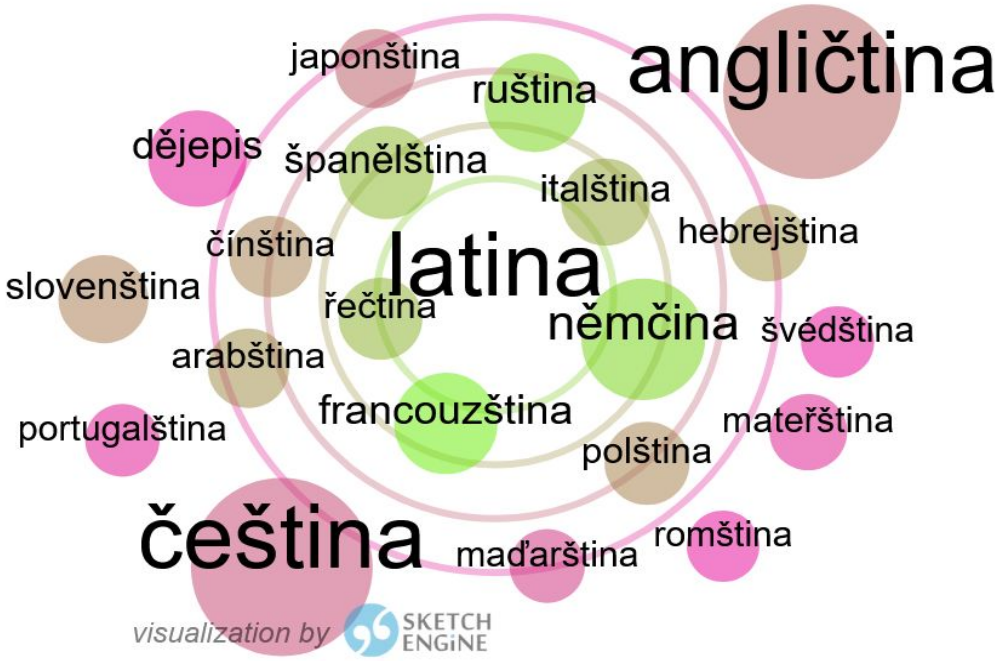
Zrada v podobě homonymie (paronymie)

jazz/tancovat/tačit

Funkce Thesaurus (zobrazení podobných slov)

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

	Word	Frequency ?	
1	francouzština	67,344	...
2	ruština	59,773	...
3	němčina	144,471	...
4	španělština	44,166	...
5	řečtina	20,932	...
6	italština	29,962	...
7	hebrejština	13,115	...
8	arabština	16,590	...
9	polština	22,629	...
10	čínština	19,423	...



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

- Společné kontexty
- Kontexty typické pro každý člen dvojice

verbs with "latina/čeština" as subject

přítelkinout	6	0	•••
šukat	6	0	•••
vytlačovat	10	8	•••
skloňovat	7	14	•••
vyučovat	13	39	•••
užívat	20	66	•••
rozlišovat	11	100	•••
znát	12	236	•••
neznat	0	82	•••
přechylovat	0	53	•••
chybět	0	321	•••
chybit	0	336	•••

"latina/čeština" and/or ...

starořečtina	79	0	•••
řečtina	1,464	42	•••
staroslověnština	60	9	•••
hebrejština	81	35	•••
italština	148	162	•••
francouzština	271	663	•••
němčina	535	2,362	•••
dějepis	69	537	•••
ruština	74	1,055	•••
polština	35	782	•••
angličtina	194	5,352	•••
slovenština	15	4,105	•••

verbs with "latina/čeština" as infinitive object

sekat	8	0	•••
učit	11	50	•••
používat	7	63	•••
přeložit	27	737	•••
vychutnat	0	6	•••
hrat	0	7	•••
znít	0	18	•••
doinstalovat	0	10	•••
odinstalovat	0	20	•••
přechylovat	0	16	•••
skloňovat	0	23	•••
překládat	0	88	•••

verbs with "latina/čeština" as genitive object

topit	20	0	•••
užívat	13	37	•••
užít	0	10	•••
vzdát	0	9	•••
všímat	0	7	•••
držet	0	49	•••
dočkat	0	54	•••
slyšet	0	13	•••
týkat	0	323	•••
bát	0	58	•••
obehrát	0	88	•••
přeložit	0	8	•••

Vyzkoušejte

- Pomocí korpus based výzkumu potvrďte/vyvraťte tvrzení, že tvar tzv. l-ového příčeští maskulina singuláru v češtině musí končit na ***–l***.
- Pomocí introspekce sestavte seznam spojení adjektivum červený+ substantivum takové, že jde o termín. Pomocí rozhraní Sketch Engine vytvořte slovní profil adjektiva ***červený*** apodívejte se, která spojení jste si vybavili a na která jste zapomněli.
- Jaká adjektiva si vybavíte, když se řekne ***stísněný***? Pomocí rozhraní Sketch Engine a funkce Thesaurus vytvořte seznam/ word cloud takových adjektiv vygenerovaných z korpusu czTenTen17 a porovne jej opět s tím, který jste získali pomocí introspekce.
- Zamyslete se, které substantivum lze rozvíjet adjektivem ***stísněný*** a které nelze rozvíjet adjektivem ***přeplněný*** a naopak. Podívejte se na **Word Sketch Difference (**nouns modified by "stísněný/přeplněný"**)** a porovnejte introspekci s daty získanými z korpusů.

Vyzkoušejte

- Definujte význam substantiva ***vařič***. Pomocí nástroje kolokace ověřte úplnost definice.
- Slovotvorný význam slov jako ***snoubič, hořčák, sněhule, voláč*** porovnejte s významem lexikálním - použijte korpus.
- Při vyhledávání v korpusech pracujeme se zadáním formálních požadavků, které hledané slovo musí splňovat. Mnohdy ovšem narážíme na to, že forma, kterou hledáme je víceznačná (problém homonymie). Tak např. slova ***datel*** a ***skladatel*** končí stejně, ale jinak nemají mnoho společného. Přesto nejsme vždy odkázáni na ruční třídění dat. Česká příjmení z l-ových příčestí (***Skácel, Přecechtěl, Snášel, ...***) mohou končit na ***el***, a přesto lze z gramatických pravidel češtiny dokázat, že nemohou být homonymní s činitelskými jmény derivovanými příponou *-tel*. Jaká omezení platí pro konsonanty, které mohou předcházet zakončení *[eě]l* českých sloves?

V násedujícím textu vidíme příklad jazykového humoru založeného na homonymii jazykových prostředků:

- Dalo se předpokládat , že po týhle exekuci už dá pokoj . „ Tak co je , Jíchová , co je . ? Kdybyste laskavě ráčila hejbnout zadkem . Nemáme času nazbyt , musíme ještě probrat Povolží , Podněstří . . . " „ Pozvrací a Poprdí . . . " ozvalo se zezadu . „ Píšu si vás , Joch ! Píšu si vás do třídní knihy ! " zaječela Ema . „ Ale to jsem nebyl já . " „ To mě nezajímá . Řekla jsem , že si vás píšu .

Pokuste se odpovědět na následující otázky:

Substantivum ***Porýní*** vykazuje jistou podobnost s verbálními substantivy typu ***poznání, podání, pokračování, …*** Pokud budeme chtít vyhledat pouze verbální substantiva, pak máme k dispozici znalosti o jejich formálních vlastnostech.

Které to jsou?

Můžeme na základě znalosti formálních vlastností verbálních substantiv substantiva ***Porýní, Pohroní, Poberouní*** vyloučit? A jak je na tom ***Pomohaní***?

Děkuji vám za pozornost