KonText a Sketch Engine – shody a rozdíly korpusových vyhledávačů v českém prostředí[1]

Klára Osolsobě

osolsobe@phil.muni.cz

Anotace

Přednáška se zaměří na praktický úvod do práce s jazykovými korpusy češtiny. Ukážeme nejdříve základní funkce nástroje pro práci s korpusy KonText a porovnáme je s analogickými funkcemi nástroje Sketch Engine. Na závěr vysvětlíme, kterými funkcemi disponuje nástroj Sketch Engine navíc oproti nástroji KonText a budeme demonstrovat, jak lze tyto funkce využít pro studium češtiny.

Úvod

Jednou z podmínek, kterou musí jazykový korpus jako zdroj dat pro studium jazyka splňovat, je v moderní době elektronické uložení a elektronická přístupnost jazykových dat, která korpus tvoří. Pro češtinu existuje v současnosti velké množství jazykových korpusů, které tuto podmínku splňují. Přestože roku 1994 vzniklo centrální pracoviště, jímž je Ústav Českého národního korpusu na filozofické fakultě Univerzity Karlovy v Praze (http://ucnk.ff.cuni.cz/cs/), dosud přetrvává v českém prostředí jistá variabilita nástrojů počítačového zpracování přirozeného jazyka (dále NLP). Nejobvyklejším základním nástrojem pro práci s jazykovým korpusem je korpusový vyhledávač (manažer), který v zásadě umožňuje klást dotazy na jazykové jednotky a zobrazovat jejich výskyt ve formátu KWIC (**k**ey **w**ord **i**n **c**ontext).

Standardní korpusový vyhledávač (manažer) disponuje dalšími funkcemi, které umožňují **ukládat** vyhledaná data, **třídit** je a **analyzovat** je, přičemž nejobvyklejším typem analýzy je **frekvenční analýza** a dále analýzy založené na složitějších **statisticky ověřených metodách** pracujících s jazykovými daty. Výsledkem jsou potom frekvenční seznamy a seznamy kandidátů na relevantní kolokace pro klíčovou jednotku (klíčové slovo).

Situace v českém prostředí

V českém prostředí (České republice) lze z uživatelského hlediska pracovat s několika podobami korpusových vyhledávačů, které se váží na akademická pracoviště. Na tomto místě pomineme vyhledávač B. Lehečky vytvořený pro Staročeskou textovou banku v rámci projektů pracoviště Ústavu pro jazyk český AV ČR (http://vokabular.ujc.cas.cz/banka.aspx?idz=STB) a budeme se věnovat patrně uživatelsky nejrozšířenějšímu vyhledávači, jímž je nástroj **KonText**. (https://kontext.korpus.cz/first_form) a dále komerčnímu produktu firmy Lexical Computing

[1] Text vznikl za podpory projektu **MUNI/A/0862/2017** Čeština v jednotě synchronie a diachronie – 2018.

Sketch Engine (https://www.sketchengine.co.uk/), který vybíráme zejména proto, že nástroj – webové rozhraní **KonText** s ním historicky souvisí.

KonText a Sketch Engine historicky

KonText je rozšířenou a graficky upravenou verzí původní aplikace **NoSketch Engine**, vyvíjenou v ÚČNK FF UK a na ÚFAL MFF UK pod licencí **GNU GPL 2**. Stejně jako **NoSketch Engine**[2] používá i KonText jako výpočetní server program Manatee.

Vyhledávač (grafické rozhraní) **KonText** slouží všem uživatelům, kteří pracují s korpusy zpřístupňovanými Ústavem Českého národního korpusu. Projekt **Český národní korpus** umožňuje pracovat s obecnými vyváženými korpusy synchronními i diachronními, psaného i mluveného jazyka. Kromě toho jsou k dispozici i různé specializované korpusy, mezi nimi i korpusy paralelní (více informací viz http://wiki.korpus.cz/doku.php/cnk:uvod). Velké synchronní psané korpusy jsou lemmatizovány a morfologicky (některé i syntakticky) označkovány.

Sketch Engine je nástroj pro výzkum fungování přirozeného jazyka určený lingvistům, lexikografům, překladatelům a učitelům i studentům. Umožňuje přístup ke čtyřem stům korpusů, které představují bohatství devadesátky jazyků (přehled a seznam viz https://www.sketchengine.co.uk/user-guide/user-manual/corpora/corpora-list/).

Pro akademickou obec (studenty, pedagogy) Masarykovy Univerzity je dostupný již delší dobu. Nově od dubna 2018 do dubna 2022 je volně dostupný pro nekomerční využití přes infrastrukturu EU ELEXIS (https://www.sketchengine.co.uk/elexis/).

Přehled základních funkcí obou nástrojů

Pokud chceme užívat jakýkoliv korpusový vyhledávač, volíme na začátku korpus, s nímž chceme pracovat. Odpověď na otázku, který z obou nástrojů, o nichž je řeč výše, je lepší, je podmíněna badatelským záměrem. Ten totiž musí vzít v úvahu i to, který korpus je vhodný k jeho uskutečnění.

Výhodou korpusů nabízených přes ÚČNK je dlouhodobá orientace pracoviště na tvorbu vyvážených korpusů češtiny a na dobrou úroveň lingvistických anotací. Výhodou nástroje Sketch Engine je pestrá nabídka korpusů různých jazyků. Se zřetelem k češtině spatřujeme jisté omezení v tom, že nabízí menší výběr korpusů než je k dispozici v ÚČNK. Nespornou výhodou nástroje SketchEngine je řada funkcí, které lze využít zejména v moderní korpusově orientované lexikografii. Některé z nich chceme ukázat, neboť je mohou využít i pokročilí studenti češtiny, a sice jako svého druhu on-line slovník.

[2] Manažer užívaný dříve pro práci s korpusy ÚČNK, který je nekomerční volně dostupnou (open-source) verzí softwarového nástroje Sketch Engine. Na rozdíl od něho nedisponuje některými funkcemi. Více: https://nlp.fi.muni.cz/trac/noske.

Shodné funkce obou nástrojů

Nadále se tedy zmíníme o těch funkcích, které jsou společné, a ukážeme na praktických příkladech, jak je lze použít.

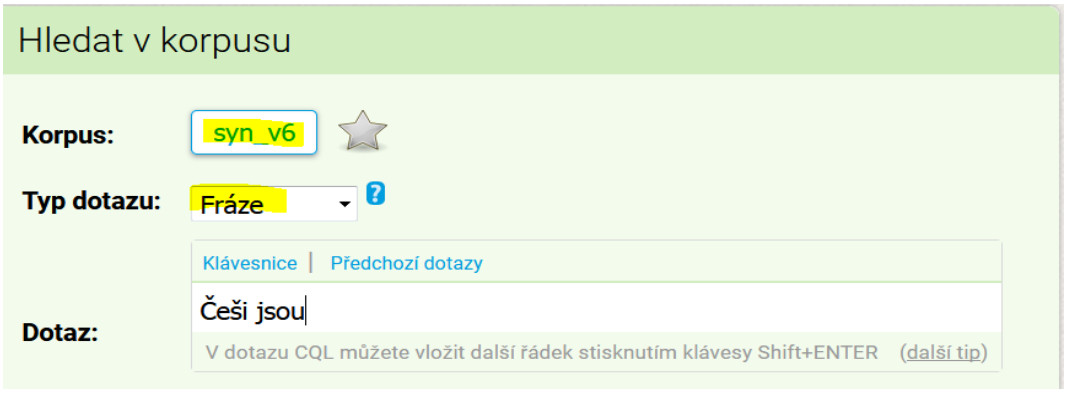
Základní funkcí každého korpusového vyhledávače je okno, do kterého lze zapsat dotaz na jazykovou jednotku, kterou chceme v korpusu najít. Můžeme se tázat na konkrétní slovní tvar (**word**), na slovní spojení (**frázi**), popřípadě můžeme konstruovat složitější dotaz za pomoci regulárních výrazů v jazyce **cql** (**c**orpus **q**uerry **l**anguage). Pokud je korpus lemmatizovaný a značkovaný, pak se lze dotazovat také na všechny tvaroslovné varianty patřící pod jeden základní tvar, tedy na systémové slovo neboli **lemma**. **Typ dotazu** je v obou vyhledávačích třeba zvolit a dodržet formální náležitosti. V obou nástrojích fungují přístupy k manuálům – uživatelským příručkám[3], v nichž jsou popsány všechny podrobnosti včetně například základních rysů jazyka **cql**.

Odpovědí na dotaz je vyhledaná jázyková jednotka ve všech výskytech ve zvoleném korpusu ve formátu KWIC (**k**ey **w**ord **i**n **c**ontext).

Mějme následující úkol. Chtěli bychom z korpusů zjistit, co říkají Češi sami o sobě. Budeme postupovat co nejjednodušeji tak, abychom přitom demonstrovali funkce obou nástrojů a zároveň porovnali výsledky na různých korpusech, které jsou přes srovnávané nástroje přístupné.

V rozhraní KonText zvolíme největší v současnosti dostupný korpus synchronní psané češtiny **syn_v6** a přes **Typ dotazu** Fráze vyhledáme všechny výskyty spojení *Češi jsou* ve zvoleném korpusu (Obrázek 1).

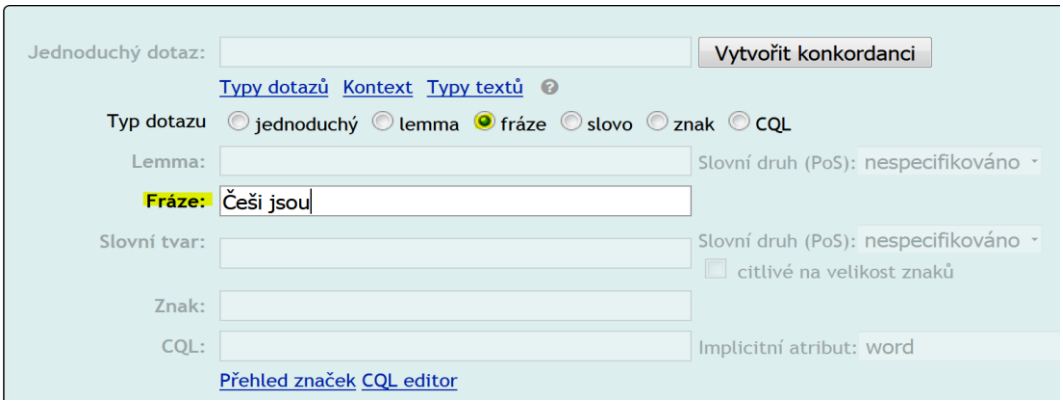
Obrázek 1



Tutéž akci zopakujeme pomocí nástroje Sketch Engine, přičemž budeme pracovat s velkým webovým korpusem **czTenTen12** (Obrázek 2).

Obrázek 2

[3] V rozhraní KonText se přes menu **Nápověda/Help → Uživatelská příručka/User Manual** se podle volby jazyka dostaneme k podrobné internetové příručce v češtině nebo angličtině. V nástroji Sketch Engine je v liště přímo odkaz **Uživatelská příručka/Help** oba odkazy vedou k manuálu v angličtině.



Výsledkem bude v obou případech konkordanční seznam (Obrázek 3 a 4).

Obrázek 3

víru v lidskost , ale i v to , že	Češi jsou	přece jen národem muzikantů a ne zlodějů , " dodaly
na dnešním losování základních skupin v Kyjevě zvláštní premiéru .	Češi jsou	totiž poprvé v samostatné historii žebříčkově nejhorší zemí , která
řekl . Třídíme nejvíce odpadu ZLATÁ MEDAILE V RECYKLACI .	Češi jsou	v třídění odpadu první v Evropě , lidé z Pardubického
ještě lépe postaveného Šmicra , přiťukává mu balon a všichni	Češi jsou	v extázi ! Tedy ve fotbalovém nebi pro puristy .
3 . Co máte a nemáte rád na Češích ?	Češi jsou	obecně velmi milí , laskaví a skromní lidé . Někdy

Obrázek 4

a neúspěchem ! A ať se to někomu líbí nebo ne , **Češi jsou** favorité . Blesk bodoval deset kategorií (5
musí Severany jednoznačně smáznout ! Forma **Češi jsou** zkrátka mazáci ! Před EURO " simulovali " že jim
tři výhry a při každé z nich nejprve prohrávali . **Češi jsou** mistři v otáčení zápasů ! Češi : 5 Dáni : 4
. Zraněný je obránce Jensen a útočník Sand . **Češi jsou** naopak zdraví jako rybičky . Češi : 5 Dáni : 3
je proti . Vyplývá to z průzkumu agentury STEM . **Češi jsou** podle průzkumu poměrně dobře informováni o

Dalšími funkcemi, jimiž běžně korpusové vyhledávače disponují, jsou **ukládání** vyhledaných dat, jejich další zpracování pomocí **filtrů**, které slouží k výběru/odstranění části dat, **frekvenční analýza** vyhledaných dat a **zobrazování** dat i metadat[4].

K čemu slouží funkce **Uložit/Save**? Pomocí této funkce můžeme vyhledaná data uložit v textovém formátu a dále s nimi pracovat off-line. Pokud hledáme v korpusech ilustraci nějakého jevu v jazyce, pak nám tato funkce postačí. Příkladem je třeba využití jazykových korpusů v pedagogické praxi.[5]

K čemu slouží funkce **Filtr/Filter**? Pomocí této funkce můžeme například vybrat část konkordance. Představme si, že se chceme (v konkardančním sezamu uvedeném výše) podívat pouze na ty doklady, kdy za klíčovým slovesným tvarem *jsou* následuje adjektivum. V rozsahu hledání definovaném jako první pozice vpravo (interval <1,1>) bezprostředně po

[4] Metadata jsou jednak informace, které identifikují zdroje vyhledaných dokladů (typ textu, autora, žánr, …), jednak lingvistické informace vložené do korpusu, přes něž lze vyhledávat na různých rovinách lingvistické abstrakce (například lemmata, slovní druhy a další gramatické významy reprezentované morfologickými značkami, tzv. tagy).

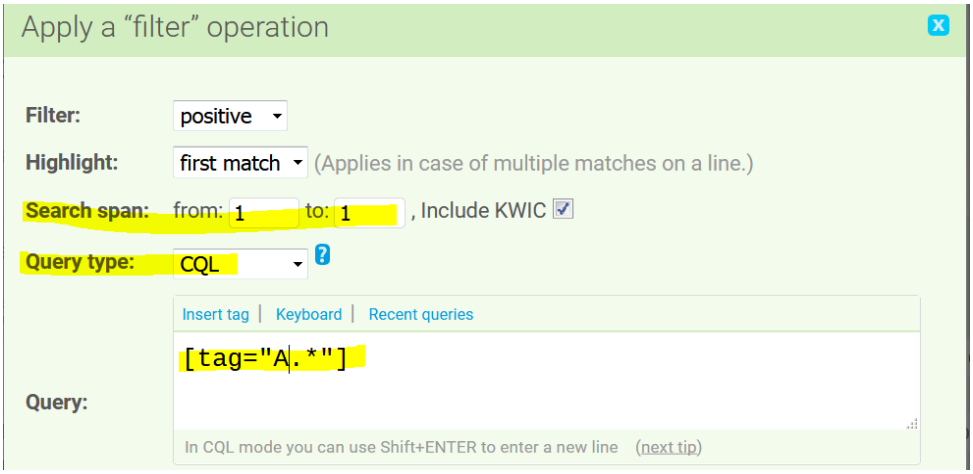
[5] Takto například pracují učitelé cizích jazyků s korpusy, když v nich hledají autentické věty jako ilustrační materiál nebo při tvorbě různých cvičení.

klíčovém slovu vyhledáme všechna adjektiva. Takto lze filtr použít, pokud je text morfologicky označkován. V jazyce cql definujeme značku podle použitého tagsetu[6].

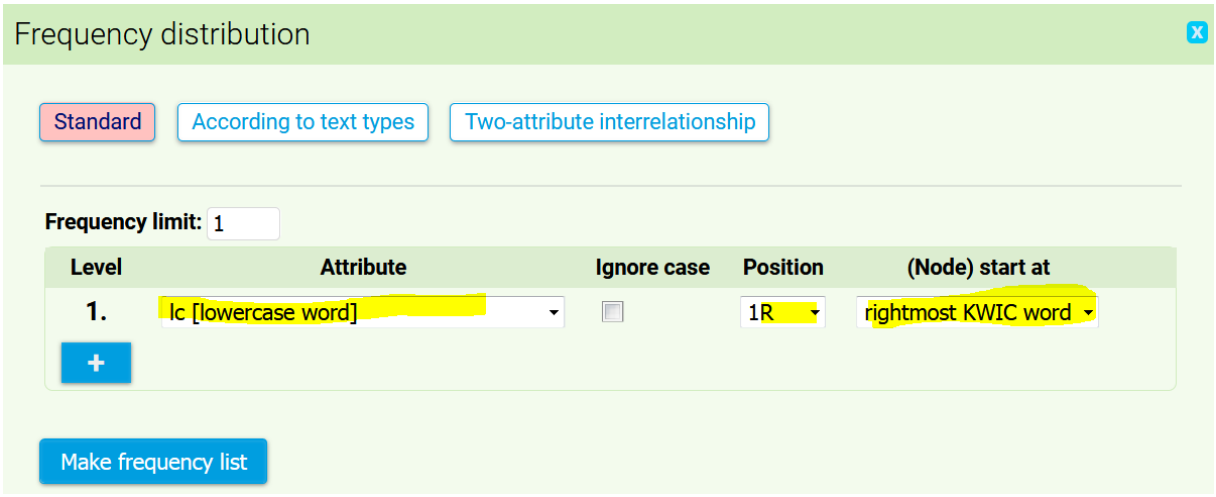
K čemu slouží funkce **Frekvence/Frequency**? Její pomocí můžeme vyhledaná data analyzovat z hlediska frekvence. Představme si, že chceme ve sledovaných korpusech zjistit frekvenci adjektiv vyskytujících se ve spojení *Češi jsou* ***<taková>***.

Na screenshotech níže vidíme, jak se pracuje v obou systémech s popsanými funkcemi a jaké výsledky získáme ze zvolených korpusů. Na první trojici screnshotů (Obrázek 5, 6, 7) vidíme zadání v rozhraní KonText a výsledek frekvenční analýzy dat z korpusu syn_v6, druhá trojice screenshotů (Obrázek 8, 9, 10) ukazuje zadání v nástroji Sketch Engine a výsledky frekvenční analýzy dat korpusu czTenTen12.

Obrázek 5

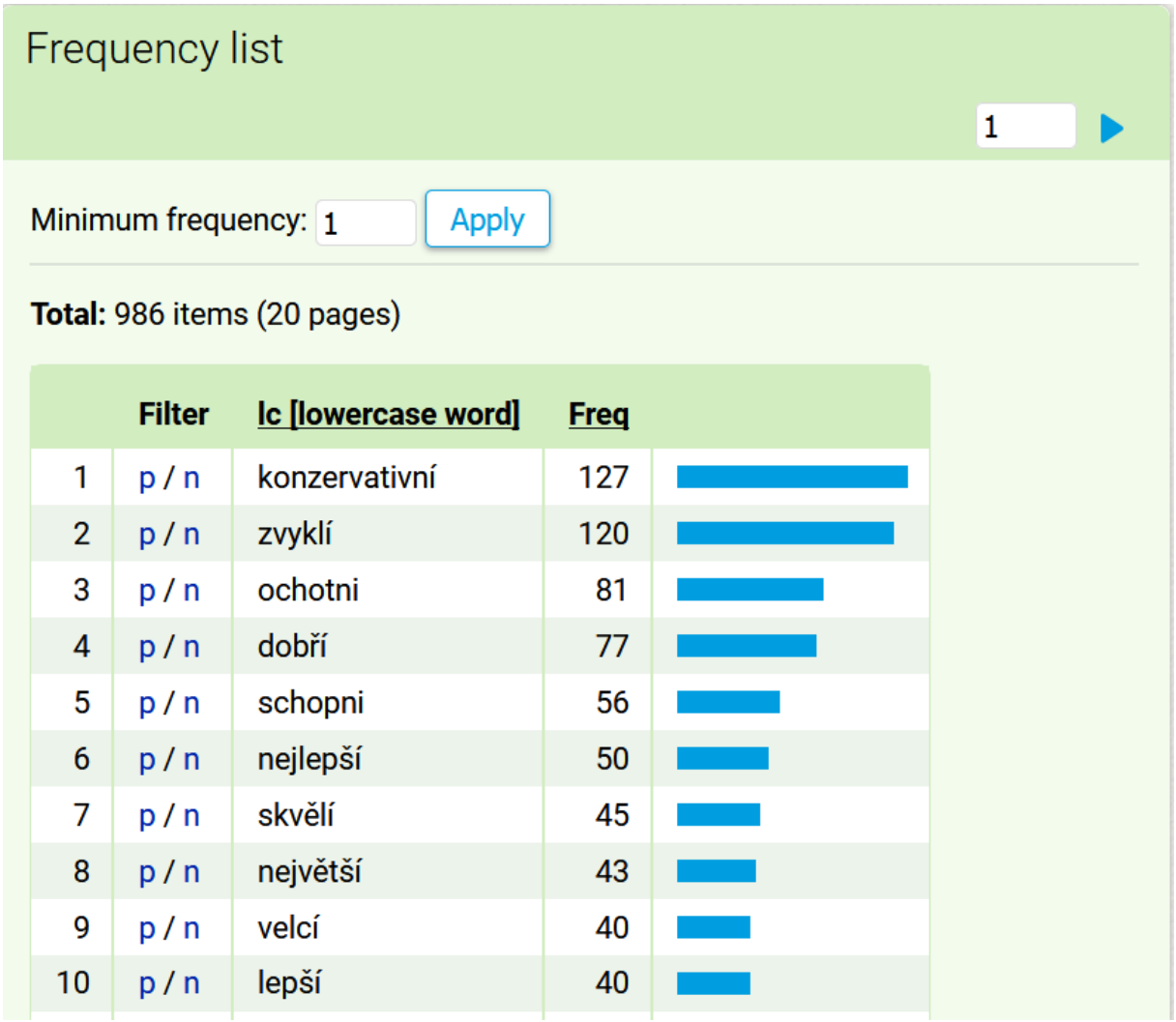


Obrázek 6



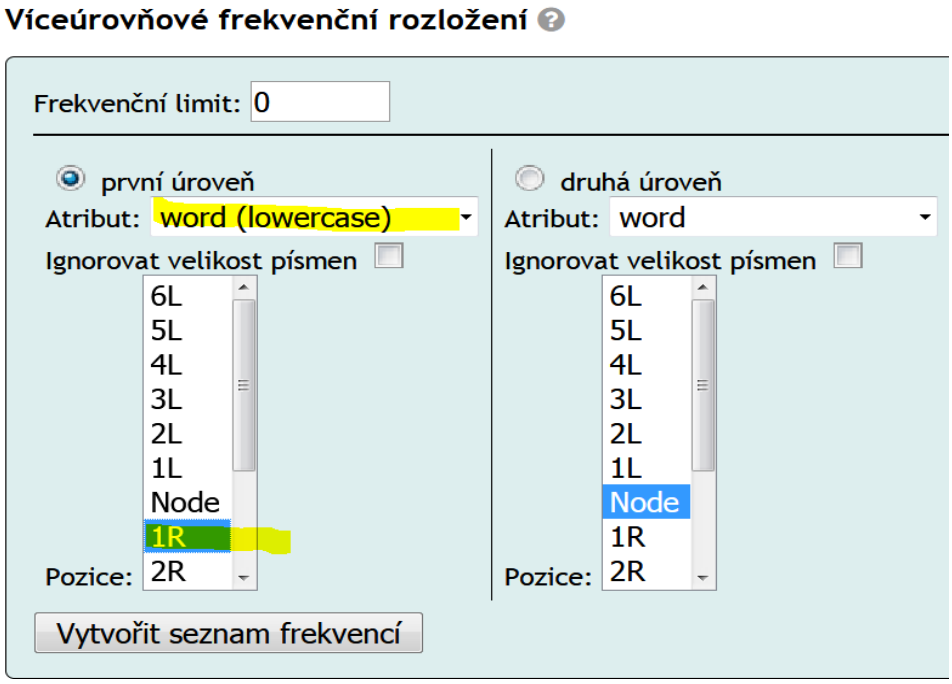
Obrázek 7

[6] V českém prostředí existuje vícero tagsetů. Korpusy přístupné přes grafické rozhraní **KonText** jsou značkovány tagsetem, který v lingvistické literatuře bývá označován jako „pražský“ /viz více zde: https://wiki.korpus.cz/doku.php/seznamy:tagy). Korpusy budované na FI MU užívají tagset označovaný jako „brněnský“ (viz https://www.sketchengine.co.uk/tagset-reference-for-czech/).

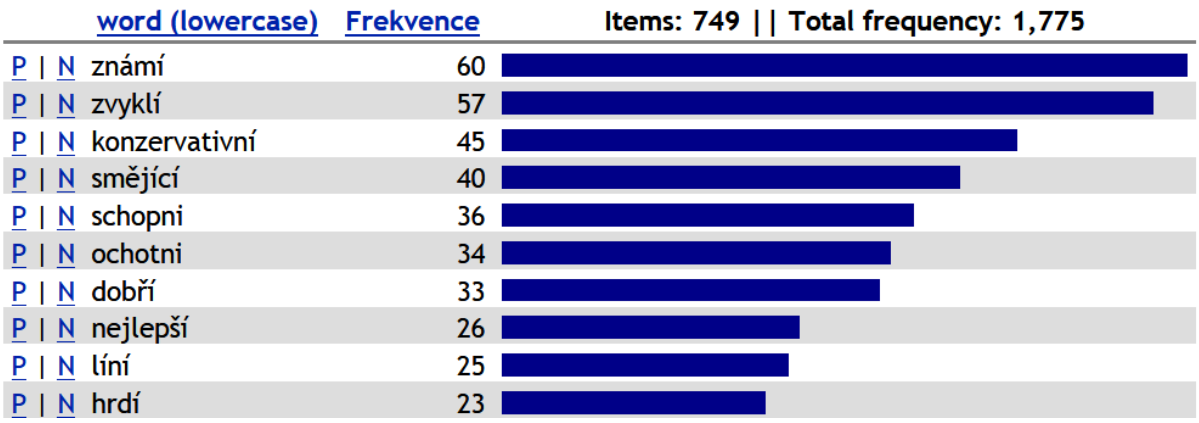


Obrázek 8

Obrázek 9

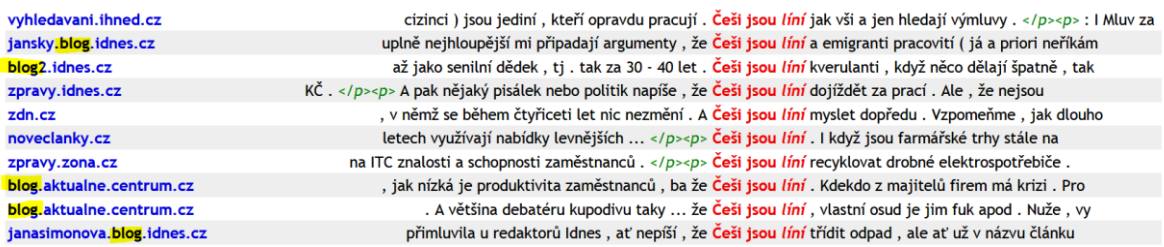


Obrázek 10



K čemu slouží funkce **Zobrazení/View**? Její pomocí měníme zobrazení konkordančního seznamu a zobrazení metainformací. Na ilustračním screenshotu (Obrázek 11) vidíme, že výskyt negativně hodnotícího adjektivum *líní*, které se jako jediné negativní adjektivum objevuje mezi prvními deseti nejfrekventovanějšími adjektivy ve sledovaném kontextu, je pozorovatelně svázán s žánrem blogu, který má z definice subjektivní charakter.

Obrázek 11

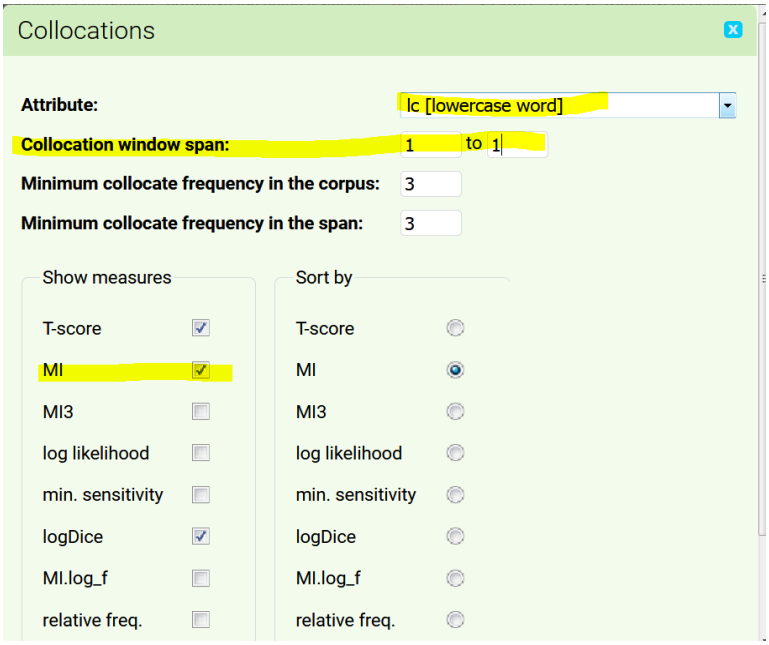


Frekvenční analýzy lze ovšem dělat také sofistikovaněji a využít k nim metody corpus driven výzkumu kolokací. Oba nástroje umožňují získat kandidáty na relevantní kolokace klíčových

slov vyhledaných v korpusech. Využívají k tomuto účelu statistické výpočty, které dávají dobré výsledky při práci s jazykovými daty[7]. Oba nástroje disponují s funkcí **Kolokace/Collocations**.

Na následujících screnshotech ukážeme přehled statisticky významných kolokací na prvním místě vpravo od klíčového spojení (fráze) *Češi jsou*. Demonstrujeme tak jak funkci obou nástrojů (Obrázek 12 a 14), tak i rozdíly ve výsledcích prosté frekvenční analýzy v porovnání se sofistikovanější analýzou relevantních kolokací (Obrázek 13 a 15). Tentokrát jsme nesledovali jen adjektiva, ale všechna slova napravo od spojení (fráze) *Češi jsou*. Výsledky jsou pouze ilustrací k fungování obou nástrojů.

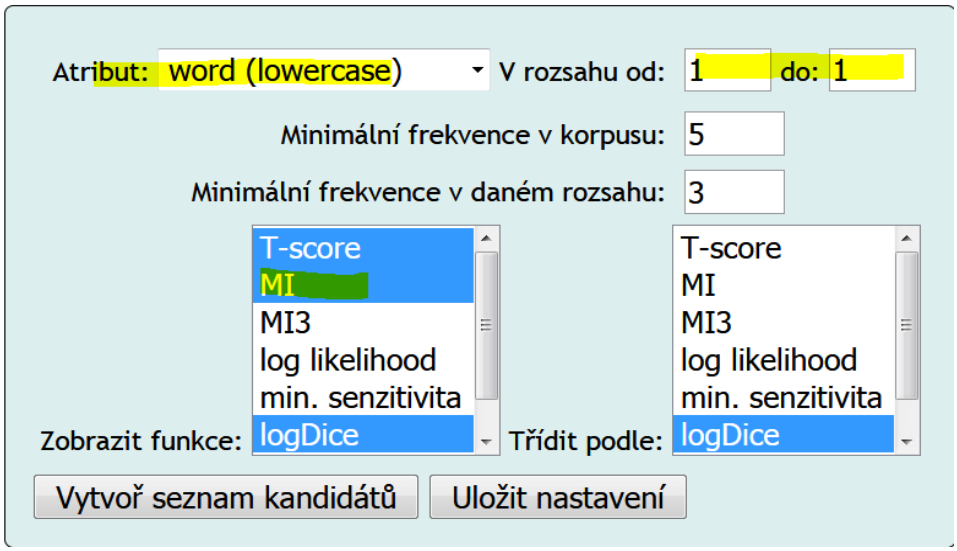
Obrázek 12



Obrázek 13

[7] Na obrázcích níže jsou kolokace seřazeny podle asociační míry MI-score (Obrázek 14) a LogDice (Obrázek 15). Viz více https://wiki.korpus.cz/doku.php/pojmy:asociacni_miry?redirect=1#dice_a_logdice.

Kandidáti na kolokace



Obrázek 14

Collocation candidates

1

	Filter	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

Obrázek 15

Kandidáti na kolokace

Strana 1 Jdi Další >

	Poučet souvýskytů	Počet kandidátů	T-score	MI	logDice
P | N **národem**	186	14,354	13.636	13.083	8.118
P | N **rasisti**	57	2,050	7.549	14.184	7.601
P | N **smějící**	40	2,646	6.323	13.305	7.003
P | N **rasisté**	28	1,818	5.290	13.332	6.610
P | N **mistři**	69	27,940	8.301	10.691	5.992
P | N **přeborníci**	16	790	3.999	13.727	5.970
P | N **národ**	230	114,572	15.154	10.392	5.947
P | N **zbabělci**	16	3,195	3.998	11.711	5.605
P | N **líní**	25	9,953	4.997	10.716	5.546
P | N **burani**	11	2,002	3.315	11.845	5.234

Funkce nástroje Sketch Engine navíc

Na závěr bychom rádi ilustrovali některé funkce, jimiž disponuje nástroj Sketch Engine navíc oproti rozhraní KontText.

První je funkce **Word Sketch/Slovní profil** sloužící k vizualizaci frekvenčně uspořádaných gramaticky definovaných relací, do kterých vstupuje klíčové slovo v daném korpusu. Na screenshotu (obrázek 16) vidíme přehled Word Sketch k lemmatu *Čech* z korpusu czTenTen12. Povšimněme si například sloves, o nichž lze předpokládat, že substantivum plní roli subjektu, a která následují příslušné substantivum.

Obrázek 16

čech *(noun)*
Czech Web 2012 (czTenTen12 v9) frekvence = 451,932 (89.14 v milionu)

a_modifier			is_subj_of			gen_2			post_verb			prec_prep		
		11.23			16.76			10.34			8.86			7.30
volyňský +	989	9.26	utratit +	422	7.39	svatopluk +	1,257	9.69	nakupovat +	254	7.19	mezi +	3,939	6.22
volyňských Čechů			jezdit +	782	7.18	Svatopluka Čecha			Češi nakupují			mezi Čechy a		
žijící +	910	7.83	Češi jezdí			praotec +	912	9.26	jezdit +	656	7.11	vůči +	468	6.08
Čechů žijících			krást +	331	6.98	praotce Čecha			Češi jezdí			vůči Čechům		
jižní +	2,815	7.78	Češi kradou			petr +	2,676	8.77	krást +	196	7.09	kromě +	395	5.07
jižní Čechy			umět +	653	6.85	Petra Čecha			Češi kradou			. Kromě Čechů		
rodilý +	389	7.61	prohrát +	373	6.72	procent +	1,914	8.71	kupovat +	264	6.88	proti +	996	4.87
rodilý Čech			Češi prohráli			procent Čechů			Češi kupují			proti Čechům		
průměrný +	1,807	7.45	nakupovat +	302	6.71	soužití +	628	8.57	věřit +	459	6.82	vedle +	171	4.63
průměrný Čech			Češi nakupují			soužití Čechů a Němců			utratit +	153	6.74	. Vedle Čechů		
etnický +	423	7.26	vyhrát +	785	6.70	třetina +	1,072	8.33	Češi utratí za			oproti +	100	4.43
etnických Čechů			Češi vyhráli			dvě třetiny Čechů			umět +	614	6.68	oproti Čechům		
východní +	1,362	7.20	bojovat +	378	6.56	ringo +	373	8.01	utrácet +	129	6.56	pro +	5,671	4.41
Východní Čechy			Češi bojovali			Františka Ringo Čecha			Češi utrácejí			pro Čechy		
hrdý +	279	7.09	věřit +	395	6.51	čtvrtina +	560	7.99	mířit +	206	6.43	včetně +	255	4.06

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.

Druhou funkcí je **Thesaurus** (zobrazení podobných slov: na základě porovnání kontextů je vytvořen seznam a vizualizace slov, která mají podobné gramaticko-lexikální kontexty). Na screenshotu (obrázek 17) jsou slova vyskytující se v podobných kontextech jako slovo *Čech*.

Obrázek 17

čech *(noun)*
Czech Web 2012 (czTenTen12 v9) frekvence = 451,932 (89.14 v milionu)

Lemma	Skóre	Frekvence
německ	0.480	252,884
američan	0.448	172,108
obyvatel	0.419	656,805
občan	0.414	818,691
student	0.412	847,292
politik	0.410	432,673
rom	0.401	175,093
rodič	0.395	949,755
národ	0.395	415,829
muž	0.394	1,616,463
podnikatel	0.389	289,755
učitel	0.389	456,974
hráč	0.388	1,580,578
člen	0.387	1,391,941
křesťan	0.385	194,918
žena	0.384	1,899,452
cizinec	0.381	151,574

Třetí funkcí je **Sketch rozdíl** (vizualizace shod a rozdílů kontextu dvojice slov). Na následujících screenshotech je srovnání frekvencí sloves stojících v postpozici za dvojicí *Čech/Němec* (Obrázek 18) a *Čech/Američan* (Obrázek 19). Sloveso v postpozici za slovními tvary lemmatu *Čech*, které se nikdy nevyskytuje v téže pozici spolu s lemmaty *Němec* nebo *Američan* je podle screenshotů ***krást***. Naopak slovesa stojící (v analyzovaném korpusu) pouze v kontextu lemmat *Němec* nebo *Američan* (nikoli *Čech*) jsou ***zaútočit*** a ***bombardovat***. Jediným společným slovesem vyskytujícím se v bezprostředním kontextu všech tří lemmat (*Čech/ Němec/ Američan*) je ***milovat***. Takto na základě statistických šetření řízených korpusovými daty lze vytvářet objektivní obraz o tom, **co na sebe prozradíme, když o sobě a druhých píšeme**.[8]

Obrázek 18

[8] Zajímavé je porovnání Sketch rozdílu dalších analogických dvojic jako *Čech/Rus*, *Čech/Rom* a *Čech/Žid*.

post_verb	40,024	11,123	0.09	0.04
krást	196	0	7.1	--
utrácet	129	0	6.6	--
mířit	206	13	6.4	3.1
kupovat	264	23	6.9	4.1
utratit	153	9	6.7	4.0
nakupovat	254	20	7.2	4.5
vypít	122	8	6.4	3.9
jezdit	656	100	7.1	4.7
věřit	459	78	6.8	4.6
umět	614	116	6.7	4.5
cestovat	130	14	6.3	4.3
milovat	279	52	6.4	4.4
říkat	275	289	4.9	5.1
okupovat	13	16	3.3	5.1
česat	12	16	3.2	5.3
útočit	15	29	3.2	5.3
nazývat	20	58	3.2	5.5
prchat	9	16	2.8	5.2
střílet	8	19	2.5	5.1
obchodovat	11	26	3.0	5.7
obsazovat	0	15	--	5.1
kšeftovat	0	12	--	5.1
bombardovat	0	13	--	5.1
zaútočit	0	18	--	5.3
ustupovat	0	26	--	5.4

Obrázek 19

post_verb	40,024	15,569	0.09	0.09
krást	196	0	7.1	--
vypít	122	0	6.4	--
mířit	206	13	6.4	3.0
jezdit	656	84	7.1	4.4
kupovat	264	27	6.9	4.2
preferovat	138	20	6.3	4.3
nakupovat	254	38	7.2	5.3
umět	614	152	6.7	4.9
pít	154	30	6.3	4.7
cestovat	130	26	6.3	4.9
utratit	153	35	6.7	5.7
milovat	279	108	6.4	5.4
důvěřovat	106	29	6.3	5.5
utrácet	129	39	6.6	6.0
věřit	459	300	6.8	6.5
slavit	75	69	5.1	5.5
spotřebovat	51	40	5.0	5.6
přiznávat	55	52	5.0	5.7
konzumovat	46	36	5.1	5.8
tvrdit	121	221	4.6	5.7
domnívat	34	38	4.5	5.7
odejít	10	54	2.4	5.5
zaútočit	0	28	--	5.6
bombardovat	0	34	--	6.1

Závěr

Ukázali jsme na ilustračních příkladech shodné rysy obou korpusových nástrojů použitelných pro práci s korpusy českého jazyka. Genetická příbuznost obou nástrojů je zřejmá. Uživatel se snadno zorientuje a může pracovat s oběma nástroji. Zvolí si jeden nebo druhý s ohledem na konkrétní korpusy, které manažery zpřístupňují. Oba nástroje mají velmi dobré instruktivní uživatelské příručky, takže je možné naučit se s nimi pracovat i bez pomoci učitele.

Pro studenty češtiny začátečníky je příznivé, že, manuály jsou dostupné i v angličtině.

Sketch Engine nabízí také nově nástroj **SkELL (Sketch Engine for Language Learning)**. Jedná se o jednoduchý nástroj pro studenty a učitele pro rychlé vyhledání užití jednotlivého slova/fráze v textech rodilých mluvčí (https://cskell.sketchengine.co.uk/run.cgi/skell). Nástroj disponuje funkcí pro hledání **příkladových vět** (**konkordancí**), **slovních profilů (Words Sketch)** a **podobných slov (Thesaurus)**. Jde o pomůcku, která může nahradit výkladový slovník.

Také manažer **KonText** disponuje oproti **Sketch Engine** dalšími funkcemi a nástroji, kterým jsme zde nevěnovali patřičnou pozornost. Pro studenty češtiny jako cizího jazyka upozorníme na závěr alespoň na nástroj **Treq** k vyhledávání překladových ekvivalentů založený na paralelních korpusech češtiny a řady dalších jazyků. Tento nástroj může posloužit jako doplněk/náhrada překladového slovníku.[9]

Bibliografie

Jan Hajič (2004): *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Vol. 1. Karolinum Charles University Press, Praha.

Milena Hnátková, Michal Křen, Pavel Procházka, Hana Skoumalová (2014): The SYN-series corpora of written Czech. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, s. 160–164. Reykjavík: ELRA. http://www.lrec-conf.org/proceedings/lrec2014/pdf/294_Paper.pdf

Adam Kilgarriff, Pavel Rychlý, Pavel Smrž, David Tugwell (2004). Itri-04-08 the sketch engine. *Information Technology*.

Adam Kilgarriff, Vojtěch Kovář, Simon Krek, Irena Srdanovič, Carole Tiberius (2010): A quantitative evaluation of word sketches. *Proceedings of the 14th EURALEX International Congress*, s. 372–379.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovvář, Jan Michelfeit, Pavel Rychlý, Vít Suchomel (2014): The Sketch Engine: ten years on. *Lexicography*, 1, s. 7–36.

Miloš Jakubíček, Adam Kilgarriff, Diana McCarthy, Pavel Rychlý (2010): Fast Syntactic Searching in Very Large Corpora for Many Languages. *PACLIC*, s. 741–747.

[9] Před nějakou dobou byl v českém tisku zveřejněn seznam slov, s jejichž významem si nevědí rady ani české děti. Víte, co znamenají slova: *bezbřehý, bytelný, balamutit, cudný, horlivý, hvozd, chmurný, chrabrý, jihnout, jímavý, jízlivý, kasat se, kazajka, klání, kloudný, komolit, konejšit, lačný, láteřit, ledabylý, lomozit, lpět, mamon, mdlý, mimoděk, nedůtklivě, nejapný, niterný, okounět, osočit, ostýchavý, otálet, otrapa, perný, pohnutka, pokoutně, ponurý, pookřát, pověra, potutelný, proradný, prchlivý, předpojatý, pýřit se, rmoutit, se, rozšafný, rusý, schlíplý, slídit, spílat, srdnatý, strádat, střenka, svérázný, svízel, sudí, šev, tklivý, tryznit, úděl, uhranout, unylý, úlisný, upejpat se, úporný, úskalí, uštěpačný, útlocitný, vesměs, vzývat, záhy, zakabonit se, zakolísat, záludný, zášť, zesinat, zevrubný, zmerčit, zpupný, ztepilý*? Zkuste se na ně podívat zde: http://treq.korpus.cz/index.php a zde: https://cskell.sketchengine.co.uk/run.cgi/skell#.

Tomáš Jelínek (2008): Nové značkování v Českém národním korpusu. In: *Naše řeč*, 91, 1,s. 13–20.

Vladimír Petkevič (2014): Problémy automatické morfologické disambiguace češtiny. In: *Naše řeč*, 97, 4, s. 194–207.

Pavel Rychlý (2008): A Lexicographer-Friendly Association Score. *Proc. 2nd Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN*, 2, s. 6–9.

Drahomíra Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, Pavel Květoň (2007): The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech. In: *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*. ACL 2007, Praha, s. 67–74.

Elektronické korpusy a korpusové nástroje

Křen, M. – Cvrček, V. – Čapka, T. – Čermáková, A. – Hnátková, M. – Chlumská, L. – Jelínek, T. – Kováříková, D. – Petkevič, V. – Procházka, P. – Skoumalová, H. – Škrabal, M. – Truneček, P. – Vondřička, P. – Zasina, A.: *Korpus SYN, verze 6 z 18. 12. 2017*. Ústav Českého národního korpusu FF UK, Praha 2017. Dostupný z WWW: http://www.korpus.cz

FI MU – *czTenTen12*. Centrum zpracování přirozeného jazyka FI MU, Brno. Dostupný z: WWW: http://ske.fi.muni.cz/bonito.

Nástroj *KonTetxt* dostupný z: WWW: https://kontext.korpus.cz/first_form.

Nástroj *Sketch Engine* dostupný z: WWW: https://www.sketchengine.co.uk/.

Nástroj SkELL dostupný z: WWW: https://www.sketchengine.co.uk/skell/.

Nástroj *Treq* dostupný z: WWW: http://treq.korpus.cz/index.php.