1. Manuál ke Sketch Engine

Sketch Engine je internetový počítačový program, který nalezneme na webových stránkách https://www.sketchengine.co.uk. Jedná se o korpusový manažer, tedy software určený k hlubší analýze textů v nejrůznějších jazycích. Primárně slouží ke zkoumání chování slov v rámci konkrétního kontextu.

1.1. Přihlášení

Sketch Engine je placená platforma. Před začátkem používání se musíme přihlásit, před úplně prvním použitím je nutné se zaregistrovat. Registraci provádíme prostřednictvím formuláře, který se zobrazí po kliknutí na *Sign up* v pravé horní části obrazovky na hlavní stránce manažeru (sketchengine.co.uk).

Obrázek 1

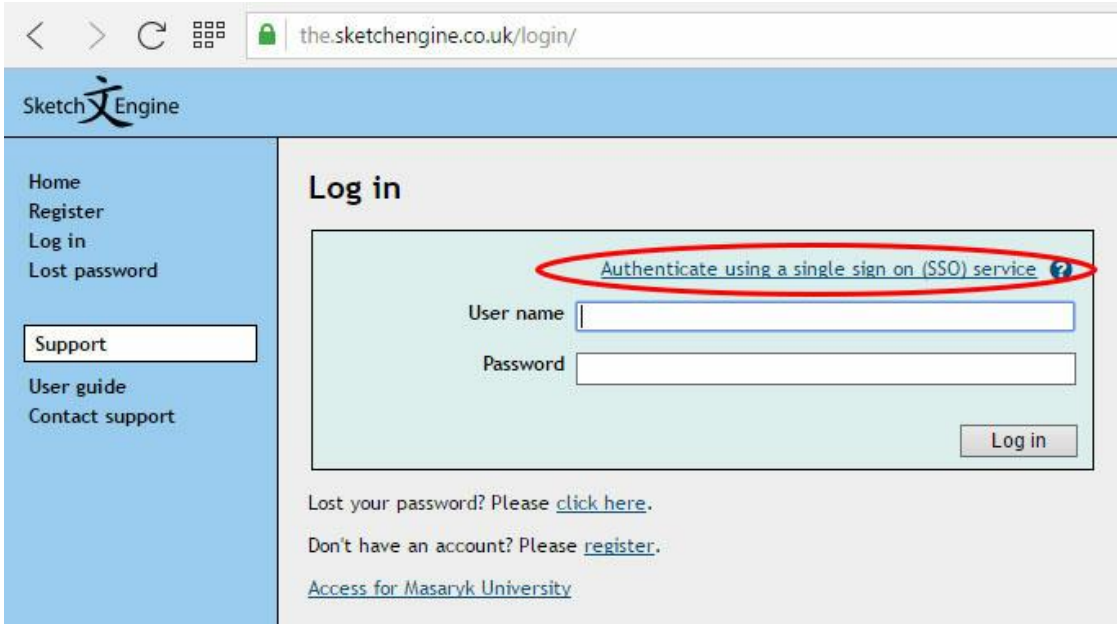
Při registraci vybíráme druh licence, prostřednictvím které chceme Sketch Engine využívat. Na výběr je:

- 30denní zkušební verze poskytovaná zdarma, ze které se po vypršení časové lhůty dá snadno přejít na plnou (placenou) verzi

Obrázek 2

- Licence pro samostatného uživatele
- Licence pro více uživatelů, které zaštiťuje jedna instituce
- Licence pro pracovníky webu (pro tuto formu registrace je vyžadován licenční klíč)

Pokud pracujeme či studujeme pod záštitou univerzity nebo jiné organizace, která si práva na používání Sketch Engine zakoupila, můžeme na hlavní stránce kliknout na *Log in* a zvolit si variantu *SSO* neboli *Single sign on* pro přístup do Sketch Engine bez nutnosti registrace do programu.



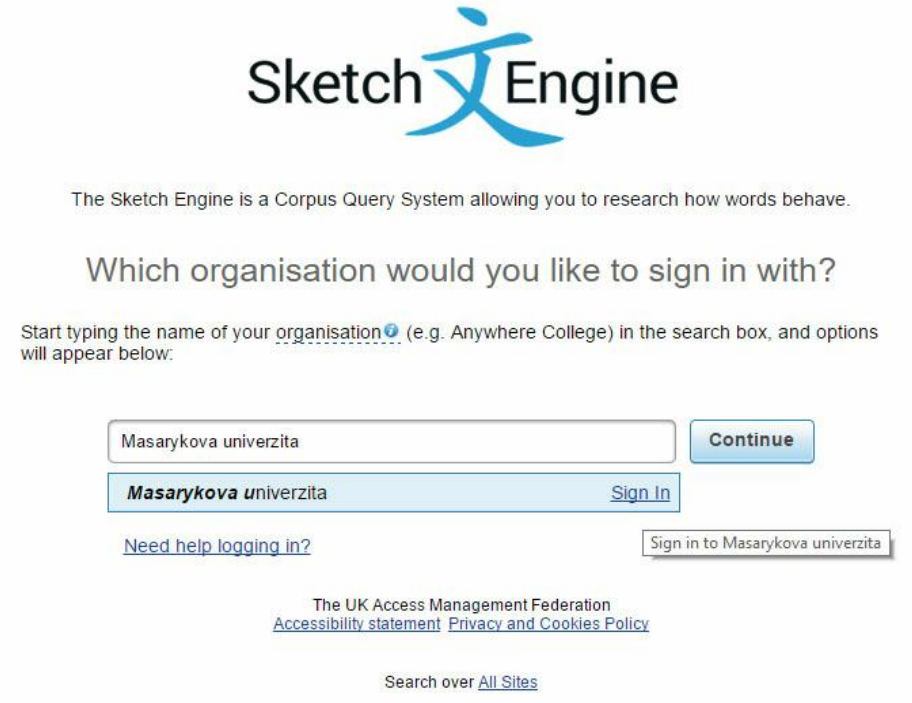
Obrázek 3

Pokud zvolíme tuto variantu přihlášení do manažeru, stačí se pak pouze přihlásit. K tomu použijeme vstupní informace potřebné k přihlášení do systémů instituce, jejímž prostřednictvím jednáme (v tomto případě je to Masarykova univerzita, proto jsou námi zadané informace ve formě UČO a sekundární heslo).

Výhody tohoto přihlášení:

- Možnost přistupovat ke všem funkcím, které Sketch Engine nabízí, zcela neomezeně
- Serveru nejsou poskytována žádná osobní data uživatelů SSO (náš přístup je tedy zcela anonymní)
- Odpadá nutnost registrace, tedy vytváření nového přístupového protokolu (šetří čas uživatele)

Obrázek 5



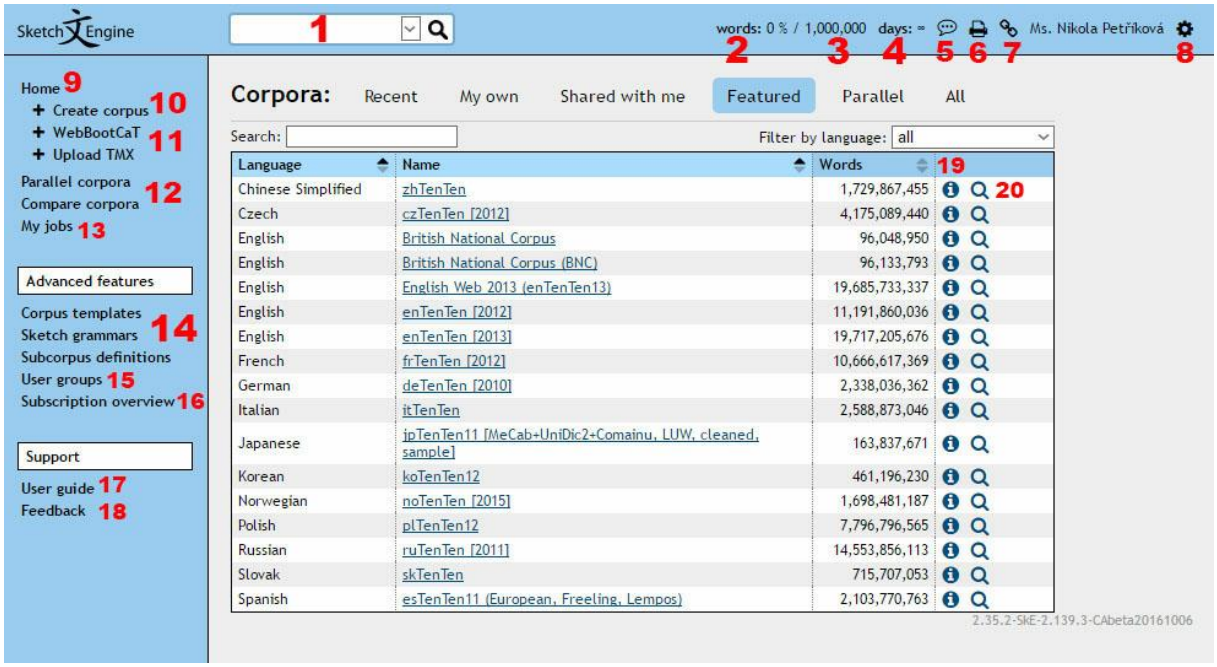
Obrázek 4

Pokud jsme si při prvním přístupu k manažeru zvolili krátkodobou zkušební verzi

programu (tzv. trial), pokračujeme po třiceti dnech v registraci a naskytnou se nám stejné možnosti jako výše popsané (viz druhý odstavec kapitoly Přihlášení). Pro uživatele, kteří nechtějí či nemohou za Sketch Engine platit, je k dispozici zdarma program NoSketchEngine (kterému se v této práci věnovat nebudeme).

1.2. Uživatelské rozhraní

1.2.1. Domovská obrazovka

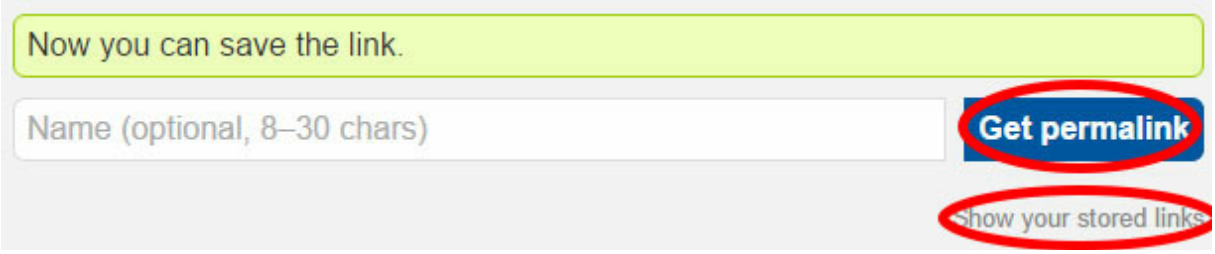


Obrázek 6

Po úspěšném přihlášení se zobrazí domovská obrazovka, která je v angličtině (tato možnost je přednastavená, a na rozdíl od uživatelského rozhraní při práci s korpusem není možné ji změnit). Na úvodní stránce je mnoho ukazatelů a informací o zobrazovaném materiálu. Jejich funkce popíšeme níže.

1. Adresní řádek pro rychlé vyhledávání konkordancí v naposledy použitém korpusu (roletkou je možné změnit korpus)
2. Procentuální množství místa, které má uživatel k dispozici při vkládání vlastního obsahu
3. Celková velikost korpusu, kterou má uživatel k dispozici při vkládání vlastního obsahu (z hlediska počtu slov)
4. Počet dní, které zbývají do vypršení předplatného
5. Hlášení o problému nebo chybě (včetně záznamu o aktuálním systémovém nastavení)
6. Tisk stránky

7. Vytvoření permanentního odkazu (tzv. permalink) kliknutím na *Get permalink*. Můžeme si odkaz osmi až třiceti znaky pojmenovat (pole *Name*). Uloží se odkaz na aktuálně zobrazovanou stránku včetně všech nastavení (použitý korpus, kritéria pro vyhledávaný řetězec, filtry atd.). Dříve uložené permanentní odkazy je možné zpětně zobrazit po kliknutí na *Show your stored links*.

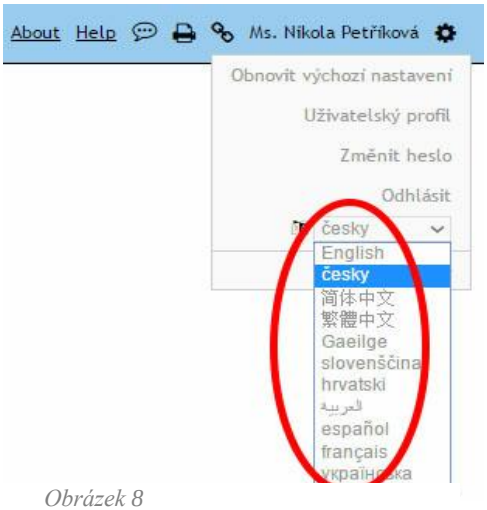


Obrázek 7

8. Nastavení uživatelského účtu (změna hesla, osobní detaily, jazyk rozhraní atd.)
9. Domovská stránka
10. Vytvoření nového/vlastního korpusu
11. Vytvoření vlastního korpusu za pomoci WebBootCaT
12. Paralelní korpus a porovnání více korpusů (viz kapitola Vstup a výstup)
13. Souhrnný výpis (historie) práce přihlášeného uživatele
14. Pokročilá nastavení podrobně rozepsaná na webu Sktech Engine v sekci Dokumentace
15. Vytvoření pracovní skupiny pro sdílení informací mezi kolegy
16. Přehled uživatelových plateb a předplatného
17. Odkaz na uživatelskou příručku v angličtině na stránkách Sketch Engine
18. Formulář na odeslání zpětné vazby
19. Informace o právě používaném korpusu
20. Prohledat korpus

1.2.2. Změna nastavení jazyka

Jazyk uživatelského rozhraní je přednastaven na angličtinu, ale na rozdíl od domovské obrazovky se dá v nastavení uživatelského účtu (viz předchozí kapitola, bod č. 8) změnit na jeden z následujících: Czech, Simplified Chinese, (Traditional) Chinese, Gaeilge, Slovene, Croatian, Arabic, Spanish, French, Ukrainian. Pro lepší názornost tohoto návodu jsme změnili nastavení jazyka na češtinu (viz obrázek vpravo).



Obrázek 8

Další možnosti nastavení uživatelského profilu

jsou intuitivní (obnovení výchozího nastavení, detaily uživatelského profilu přihlášeného uživatele, změna hesla a odhlášení ze Sketch Engine).

V případě, že si uživatel omylem nastavil jiný jazyk rozhraní, než kterému rozumí a není si jist, jak vrátit jazyk rozhraní do výchozí angličtiny, stačí použít následující odkaz: https://the.sketchengine.co.uk/bonito/run.cgi/save_global_attrs?attrs2save=uilang;uilang=en

1.3. Vstup a výstup

Zdrojem dat pro práci manažeru (tzv. vstup) může být kterýkoli z korpusů nacházejících se v jeho databázi, ty se zobrazí po kliknutí na *All* na hlavní stránce manažeru. Zde můžeme filtrovat zobrazené korpusy podle jazyka (výběrem v poli *Filter by language*).

Obrázek 9

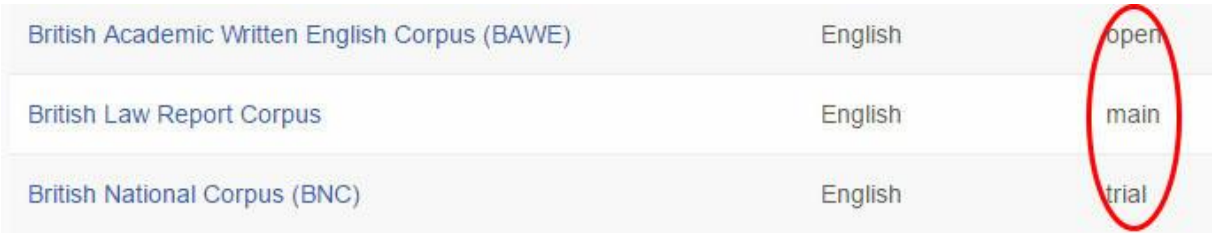
Dále může být vstupem korpus, který sami vytvoříme a do databáze vložíme (*My own*), nebo korpus, který s námi někdo sdílí (*Shared with me*).

Obrázek 10

Záložka *Recent* zobrazuje nedávno použité korpusy, *Featured* zase takové korpusy, které nám samotný Sketch Engine doporučuje k použití a *Parallel* vypíše všechny dvojjazyčné korpusy v databázi.

V databázi tohoto korpusového manažeru nalezneme mnoho korpusů v nejrůznějších jazycích, mezi nimiž je zastoupena například i maorština, telugština nebo velština. Přesně jich je v současné době 287, a to včetně 102 zkušebních (trial) verzí korpusů (tj. korpusy pro platící

uživatele i uživatele operující se zkušební verzí programu) a 4 korpusů označených jako *open* (tj. dostupné všem uživatelům i bez nutnosti registrace).



Obrázek 11

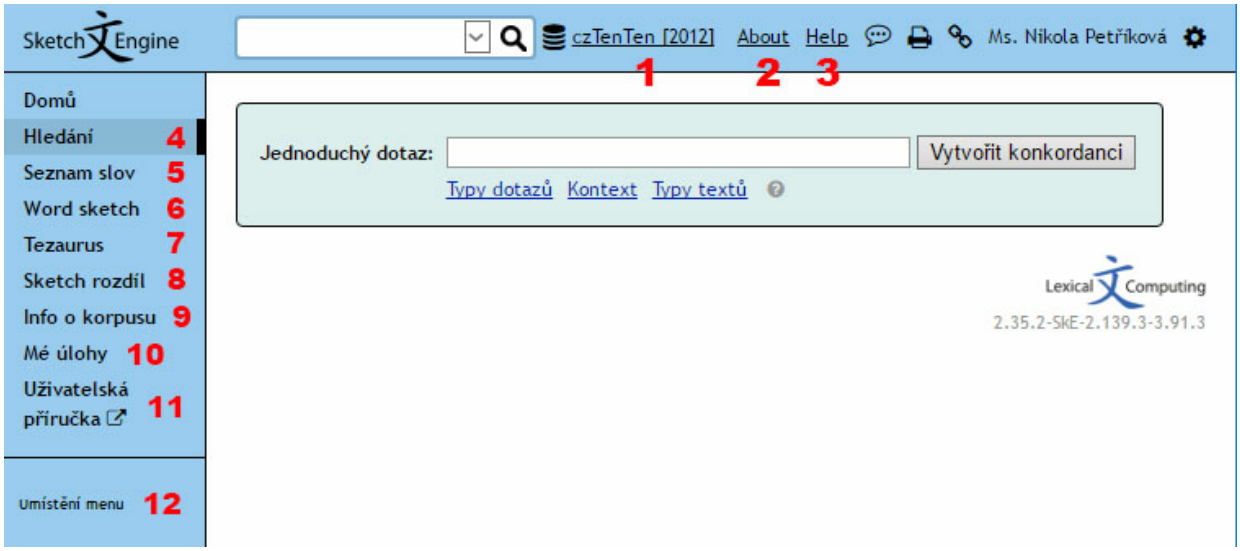
Nejčastěji využívanými výstupy programu jsou:

- *Konkordance* neboli přehled všech výskytů daného dotazu v námi zvoleném korpusu včetně textového kontextu
- *Slovní profily (Word Sketch)*, což je jednostránkový souhrn informací o gramatickém chování daného slova (dotazu) z hlediska jeho kontextu

1.4. Vyhledávání v korpusu

Tato obrazovka umožňuje vyhledávat záznamy v korpusu. Zobrazí se poté, co si uživatel vybere korpus, ve kterém chce vyhledávat, a zadá požadavek. Možnosti dostupné v této nabídce závisí na zvoleném korpusu a jeho vlastnostech a mění se podle toho, jaký jsme si zvolili výsledek vyhledávání.

1. Název aktuálně používaného korpusu
2. Odkaz na webové stránky Sketch Engine
3. Odkaz na uživatelskou příručku v angličtině na stránkách Sketch Engine
4. Popis chování zadaného slova v rámci kontextu



Obrázek 12

5. Vyhledávání konkordancí za pomoci regulárních výrazů

6. Popis chování zadaného slova z hlediska gramatiky a vazeb
7. Seznam synonym k zadanému výrazu s četností jejich výskytu
8. Porovnání rozdílů mezi dvěma zadanými výrazy z hlediska kontextu i použití
9. Informace o korpusu, ve kterém aktuálně vyhledáváme
10. Souhrnný výpis (historie) práce přihlášeného uživatele
11. Odkaz na uživatelskou příručku v angličtině na stránkách Sketch Engine
12. Odkaz pro změnu umístění menu na stránce (horizontální nebo vertikální)

1.5. Funkce Sketch Engine

1.5.1. Který korpus použít

Před začátkem používání Sketch Engine je nutné vybrat si korpus, ve kterém chceme pracovat. Pro uživatele, kteří nemají s používáním programu předchozí zkušenosti, může být problematické vybrat si vhodný korpus. V této kapitole si představíme dva základní typy korpusů (jednojazyčné a vícejazyčné) a vysvětlíme si, jaký je mezi nimi rozdíl. Konkrétní informace k jednotlivým korpusům nalezneme po kliknutí na *i* v kroužku.

Obrázek 13

1.5.1.1. Featured/All

Sketch Engine ve výchozím nastavení zobrazí korpusy v záložce *Featured*, které byly vybrány jako doporučené k použití. Pokud se pod touto záložkou nezobrazují žádné korpusy, zvolíme záložku *All* (1), nastavíme jazykový filtr na češtinu (2) a šipkou dolů seřadíme korpusy v odstavci *Words* (3) podle velikosti. Vybereme první zobrazený korpus (4), tedy z dostupných

Obrázek 14

největší. Jedná se o korpus czTenTen [2012], který se pro práci v začátcích se hodí nejlépe, jelikož obsahuje nejvíce dat.

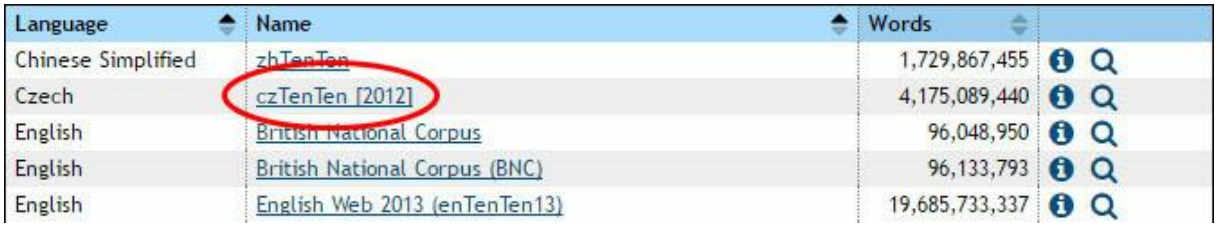
1.5.1.2. TenTen

Korpusy typu TenTen (název označuje jejich typickou velikost: 10^{10} slov) jsou jednojazyčné a velmi všestranné. Jedná se novou generaci webových korpusů a jejich hlavní

výhodou je množství obsažených dat a jejich relevance. Data, která jsou stahována z internetu, podléhají před zanesením do korpusu přísnému třídění a jsou zbavena všech netextových prvků, jakými jsou například navigační menu, právní texty, malý tisk nebo duplicitní texty. Zároveň jsou filtrovány texty, které jsou příliš krátké nebo z jiného důvodu nevhodné pro použití v korpusu.

1.5.1.3. Parallel

Paralelní korpusy jsou vícejazyčné, tedy obsahují stejný text ve více jazycích. Pro práci



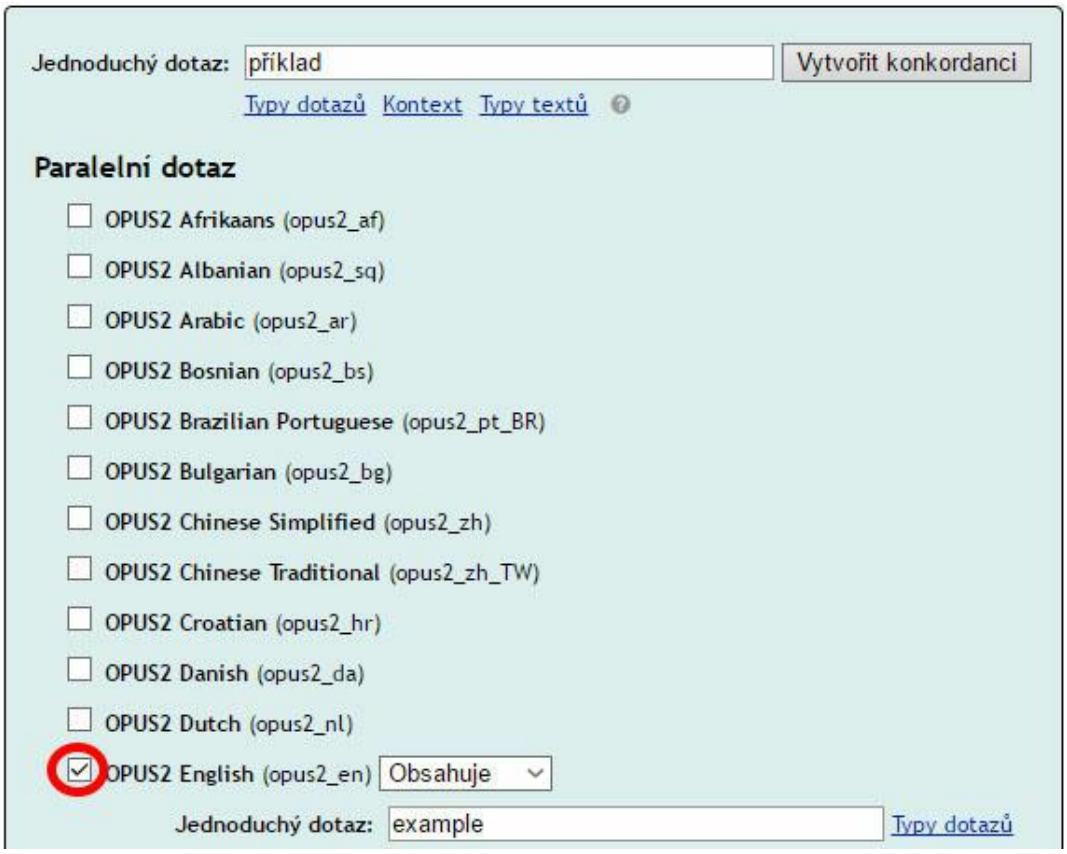
Language	Name	Words
Chinese Simplified	zhTenTen	1,729,867,455
Czech	czTenTen [2012]	4,175,089,440
English	British National Corpus	96,048,950
English	British National Corpus (BNC)	96,133,793
English	English Web 2013 (enTenTen13)	19,685,733,337

Obrázek 15

s paralelním korpusem je nejprve potřeba si na domovské stránce zvolit záložku *Parallel* a vybrat korpus v prvním jazyce. Při definování dotazu pak vybereme korpus nebo korpusy v jiném jazyce.

1.5.1.4. OPUS

Pro první práci s paralelními korpusy se doporučuje použít korpusy typu OPUS, které jsou tvořeny texty staženými z webu a následně přeloženými do ostatních jazyků, takže mají nejširší zaměření. V současné době jsou k dispozici v největším počtu jazyků.



Obrázek 16

EUR-Lex

Tyto korpusy tvoří přeložené dokumenty Evropské unie a jsou k dispozici ve 24 úředních jazycích EU. Tato dokumentace pokrývá velké množství témat, a proto se korpusy EUR-Lex hodí pro obecné zkoumání formální stránky jazyků jako takových. Dále jsou vhodným zdrojem konkrétních informací z oblastí definovaných v evropských dokumentech.

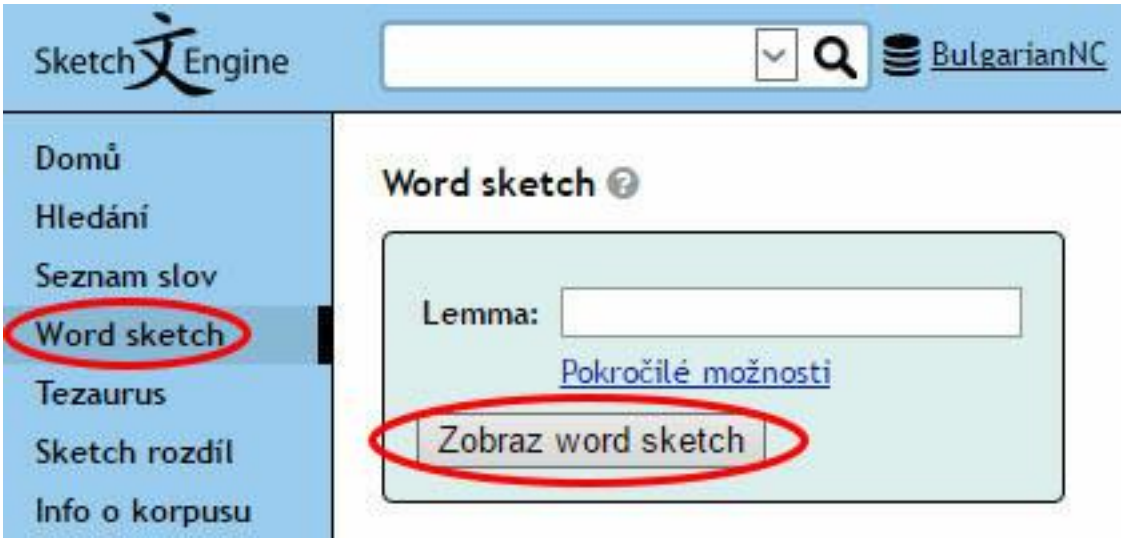
1.5.1.5. EUROPARL

Tyto korpusy jsou tvořeny daty ze záznamů jednání Evropského parlamentu. Jsou dostupné v 21 jazycích a jsou vhodným zdrojem při zkoumání témat probíraných Evropským parlamentem a obecně jako prezentace formálního užití jazyka. Nehodí se pro práci s informacemi z oblastí, které nejsou běžně předmětem těchto diskusí.

1.6. Word Sketch

Charakteristickým rysem Sketch Engine je právě Word Sketch. Je ideální pomůckou, pokud chceme zjistit, jak se které slovo chová v rámci běžného využití v jazyce. Funkce Word Sketch soustřeďuje informace o chování slova nebo fráze v kontextu, a to na základě mnoha milionů příkladů z praxe. Jejím výstupem je jednostránkový výpis slovních spojení obsahujících námi zadaný výraz, s odkazy na každé konkrétní použití. Funkce Word Sketch není dostupná pro všechny korpusy ve Sketch Engine.

Pro použití Word Sketch si vybereme libovolný korpus z nabídky kliknutím na jeho název. Pokud je u korpusu dostupná funkce Word Sketch, zobrazí se v nabídce v levém menu. Požadovaný výraz vepíšeme do kolonky *Lemma* (lemma = základní podoba slova či fráze) a klikneme na *Zobraz word sketch*.



Obrázek 17

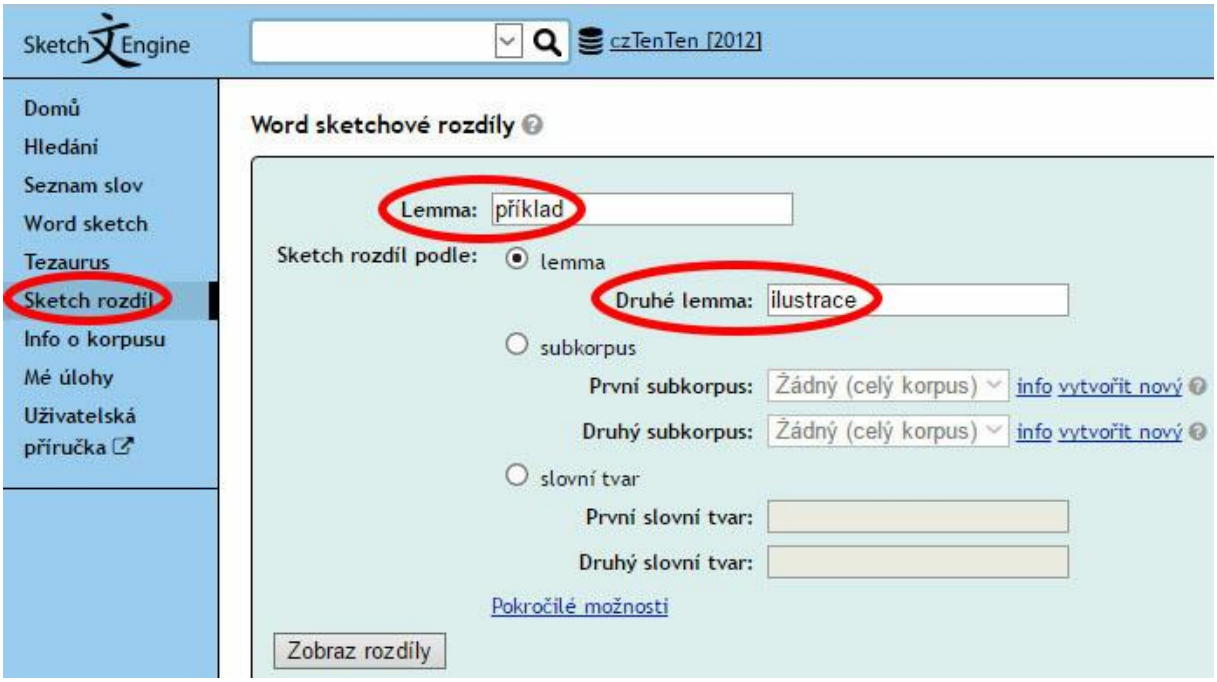
Výsledek programu se zobrazí během několika vteřin a v tomto jednostránkovém výpisu (po vyhledání lemmatu *„příklad“*) nalezneme:

1. Počet užití hledaného výrazu v celém korpusu
2. Druh vztahu mezi hledaným výrazem a kolokací (tj. spojení více slov, které spolu souvisí gramaticky i sémanticky a vytváří víceslovné pojmenování)
3. Počet užití slova v konkrétní kolokaci, po kliknutí na odkaz se zobrazí příklady konkordancí v konkrétním kontextu z příkladů v korpusu
4. Tučně jsou vypsány fráze sestávající z námi zadaného výrazu; malé plus na pravé straně ukazuje, že se modifikátor nalézá před hledaným výrazem (po kliknutí se zobrazí výpis výsledků při hledání kolokací, např. konkrétní příklad)

Obrázek 18

1.7. Sketch rozdíl

Sketch rozdíl (Sketch difference/Sketch diff) je rozšíření funkce Word Sketch a umožňuje pozorovat dva pojmy současně a porovnávat jejich chování. Hodí se zejména pro zkoumání užití synonym a antonym v jazykovém kontextu.

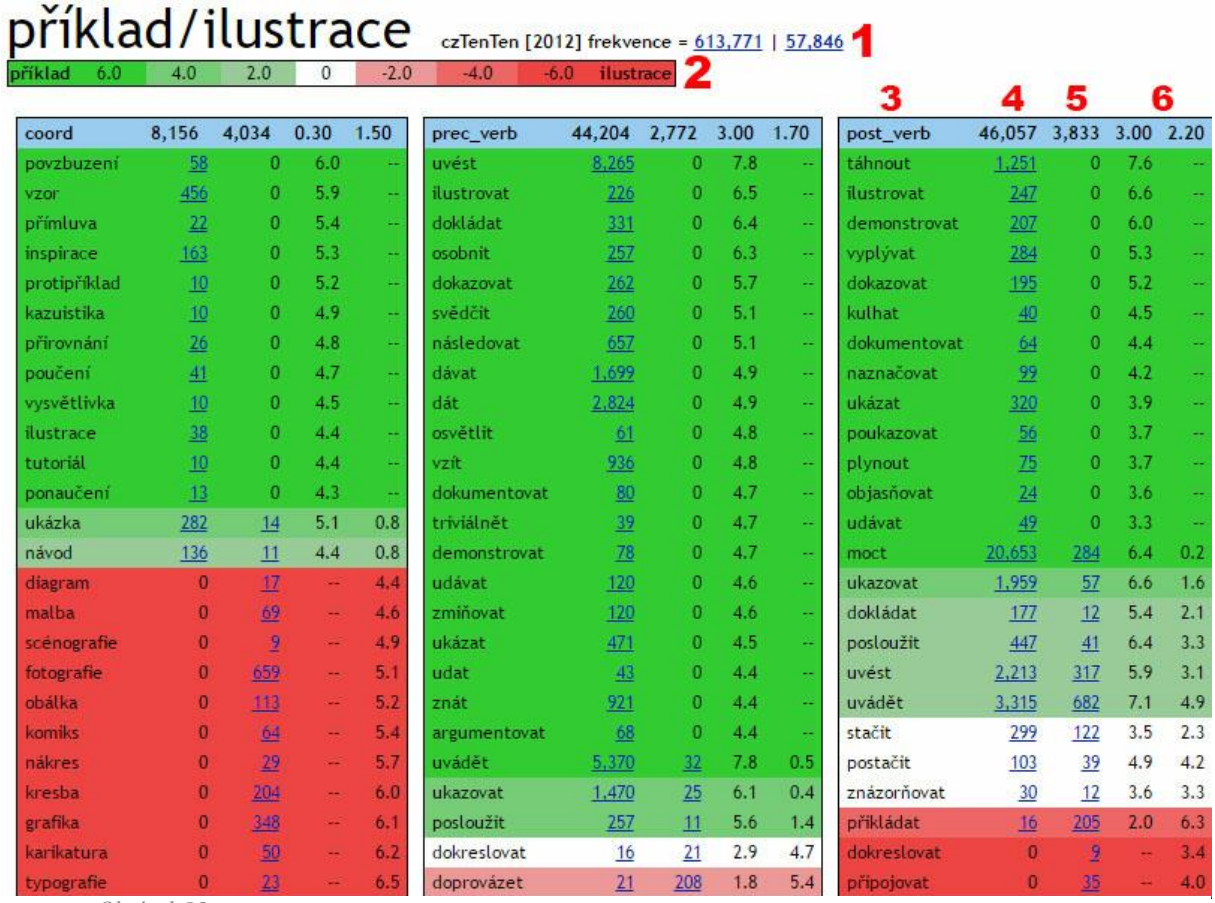


Obrázek 19

Sketch rozdíl vygenerujeme tak, že klikneme na *Sketch rozdíl* v levém menu, vepíšeme jedno slovo do kolonky *Lemma* a další do kolonky *Druhé lemma* a klikneme na *Zobraz rozdíly*.

Výsledkem je tabulka, ve které je každému lemmatu přiřazena barva (jednomu zelená, druhému červená). Červeně označené kolokace mají tendenci se pojit s červeně označeným lemmatem a zelené kolokace se zeleným. Čím sytější je barevný odstín, tím silnější je slovní spojení. Bíle označené kolokace se mohou kombinovat s oběma lemmaty.

1. Vyhledávaná slova, název korpusu a frekvence použití obou pojmů v rámci korpusu
2. Skóre obou pojmů vyobrazené v barevných odstínech



příklad/ilustrace czTenTen [2012] frekvence = 613,771 | 57,846 **1**

příklad 6.0 4.0 2.0 0 -2.0 -4.0 -6.0 ilustrace **2**

3 **4** **5** **6**

coord	8,156	4,034	0.30	1.50
povzbuzení	58	0	6.0	--
vzor	456	0	5.9	--
přímluva	22	0	5.4	--
inspirace	163	0	5.3	--
protipříklad	10	0	5.2	--
kazuistika	10	0	4.9	--
přirovnání	26	0	4.8	--
poučení	41	0	4.7	--
vysvětlivka	10	0	4.5	--
ilustrace	38	0	4.4	--
tutoriál	10	0	4.4	--
ponaučení	13	0	4.3	--
ukázka	282	14	5.1	0.8
návod	136	11	4.4	0.8
diagram	0	17	--	4.4
malba	0	69	--	4.6
scénografie	0	9	--	4.9
fotografie	0	659	--	5.1
obálka	0	113	--	5.2
komiks	0	64	--	5.4
nákres	0	29	--	5.7
kresba	0	204	--	6.0
grafika	0	348	--	6.1
karikatura	0	50	--	6.2
typografie	0	23	--	6.5

prec_verb	44,204	2,772	3.00	1.70
uvést	8,265	0	7.8	--
ilustrovat	226	0	6.5	--
dokládat	331	0	6.4	--
osobnit	257	0	6.3	--
dokazovat	262	0	5.7	--
svědčit	260	0	5.1	--
následovat	657	0	5.1	--
dávat	1,699	0	4.9	--
dát	2,824	0	4.9	--
osvětlit	61	0	4.8	--
vzít	936	0	4.8	--
dokumentovat	80	0	4.7	--
triviálnět	39	0	4.7	--
demonstrovat	78	0	4.7	--
udávat	120	0	4.6	--
zmiňovat	120	0	4.6	--
ukázat	471	0	4.5	--
udat	43	0	4.4	--
znát	921	0	4.4	--
argumentovat	68	0	4.4	--
uvádět	5,370	32	7.8	0.5
ukazovat	1,470	25	6.1	0.4
posloužit	257	11	5.6	1.4
dokreslovat	16	21	2.9	4.7
doprovázet	21	208	1.8	5.4

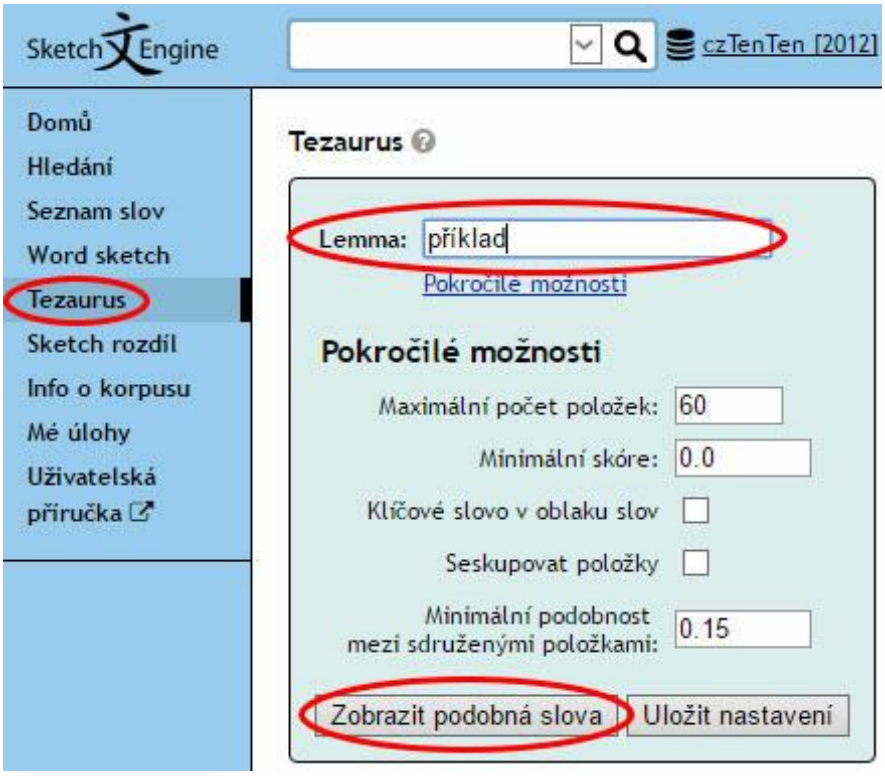
post_verb	46,057	3,833	3.00	2.20
táhnout	1,251	0	7.6	--
ilustrovat	247	0	6.6	--
demonstrovat	207	0	6.0	--
vyplývat	284	0	5.3	--
dokazovat	195	0	5.2	--
kulhat	40	0	4.5	--
dokumentovat	64	0	4.4	--
naznačovat	99	0	4.2	--
ukázat	320	0	3.9	--
poukazovat	56	0	3.7	--
plynout	75	0	3.7	--
objasňovat	24	0	3.6	--
udávat	49	0	3.3	--
moct	20,653	284	6.4	0.2
ukazovat	1,959	57	6.6	1.6
dokládat	177	12	5.4	2.1
posloužit	447	41	6.4	3.3
uvést	2,213	317	5.9	3.1
uvádět	3,315	682	7.1	4.9
stačit	299	122	3.5	2.3
postačit	103	39	4.9	4.2
znázorňovat	30	12	3.6	3.3
přikládat	16	205	2.0	6.3
dokreslovat	0	9	--	3.4
připojovat	0	35	--	4.0

Obrázek 20

3. Název gramatické kategorie slova, které se k vyhledávaným pojmům váže, např. *post_verb* znamená, že se námi zadaný pojem váže se za ním následujícím slovesem (v tomto případě: *„příklady táhnou“*)
4. Frekvence použití prvního lemmatu v kolokaci s uvedeným výrazem (detail s konkrétními příklady po kliknutí na odkaz), např. *„příklad stačí“*
5. Frekvence použití druhého lemmatu v kolokaci s uvedeným výrazem (viz výše), např. *„ilustrace stačí“*
6. Informace pro vývojáře

1.8. Tezaurus

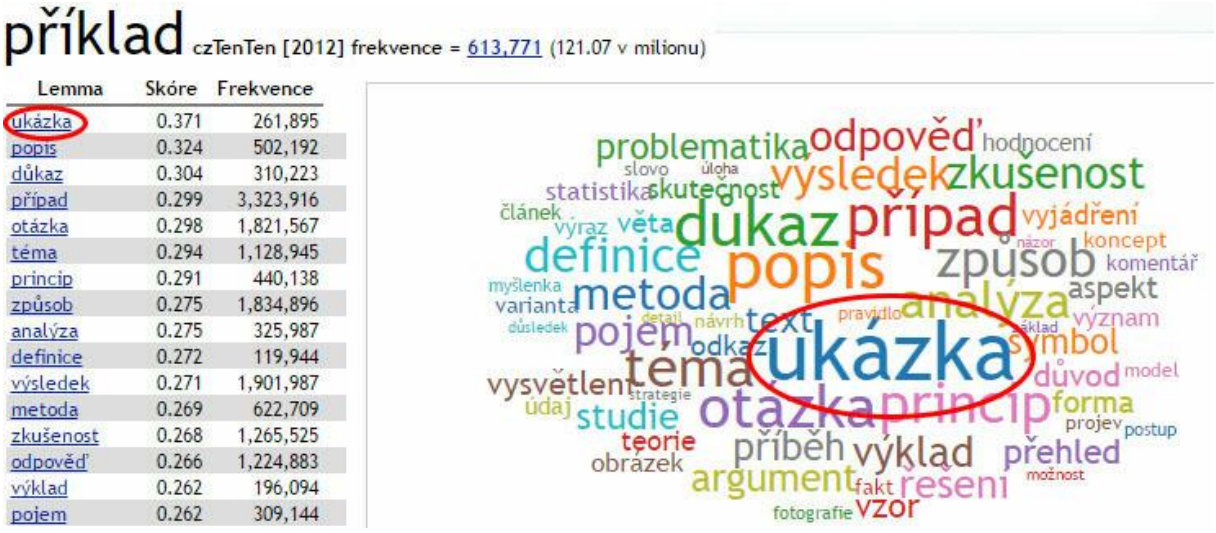
Na rozdíl od klasických tezaurů (tj. seznamů synonym) s omezeným pokrytím je ten ve Sketch Engine automaticky generován za pomoci propracovaných algoritmů analyzujících miliardy textových dat v korpusech. To znamená, že tezaurus může být vygenerován pro téměř jakékoliv slovo (za předpokladu, že korpus je dostatečně velký).



Pro použití této funkce programu Sketch Engine klikneme na *Tezaurus* v levém menu, a poté si vybereme libovolný korpus z nabídky kliknutím na jeho název. Požadovaný výraz vepíšeme do kolonky *Lemma* a klikneme na *Zobrazit podobná slova*.

Obrázek 21

Informace o synonymech k hledanému výrazu vygenerovaných modulem Word Sketch získáme po kliknutí na modrý odkaz v levém sloupci nebo na konkrétní slovo v interaktivním obrázku.



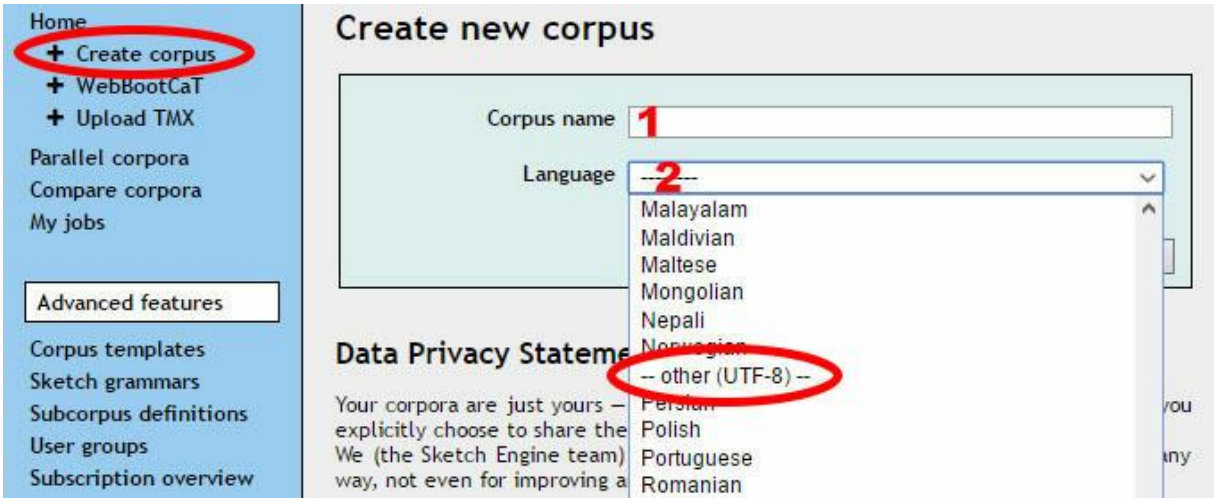
Obrázek 22

1.9. Vytvoření vlastního korpusu

1.9.1. Create corpus

Pokud chceme vytvořit nový korpus ze složky v počítači, klikneme v levém menu na domovské obrazovce na záložku *Create corpus*. Do následujícího formuláře zaneseme tyto údaje (poté klikneme na *Next >*):

1. název korpusu, podle kterého bude korpus možno vyhledat a pracovat s ním (unikátní alfanumerické ID korpusu bude vygenerováno automaticky)
2. jazyk korpusu (Sketch Engine bude s korpusem zacházet podle příslušných nastavení pro češtinu nebo jiný jazyk); pokud jazyk našeho korpusu není mezi zobrazenými možnostmi, zaškrtneme *other (UTF-8)*



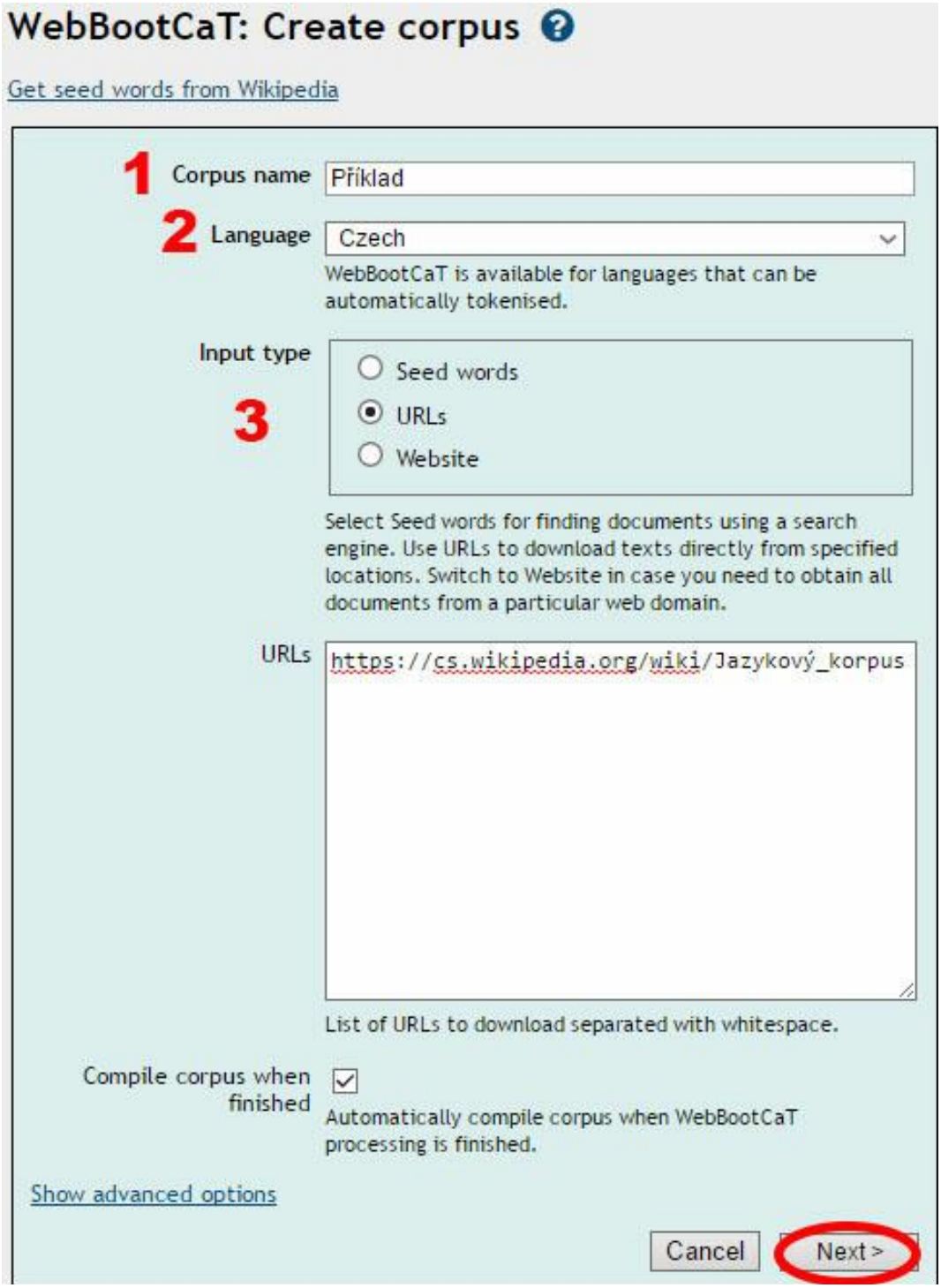
Obrázek 23

1.9.2. WebBootCaT

Tato funkce je ideální volbou, pokud chceme vytvořit korpus na základě dat stažených z webu. Aby bylo dosaženo maximální kvality extrahovaného materiálu, jsou tato data automaticky vyčištěna (tzn. zbavena duplikátů, spamu a netextových prvků).

V levém menu na hlavní obrazovce klikneme na záložku WebBootCaT. Do následujícího formuláře zaneseme tyto údaje:

1. název korpusu, podle kterého bude korpus možno vyhledat a pracovat s ním



Obrázek 24

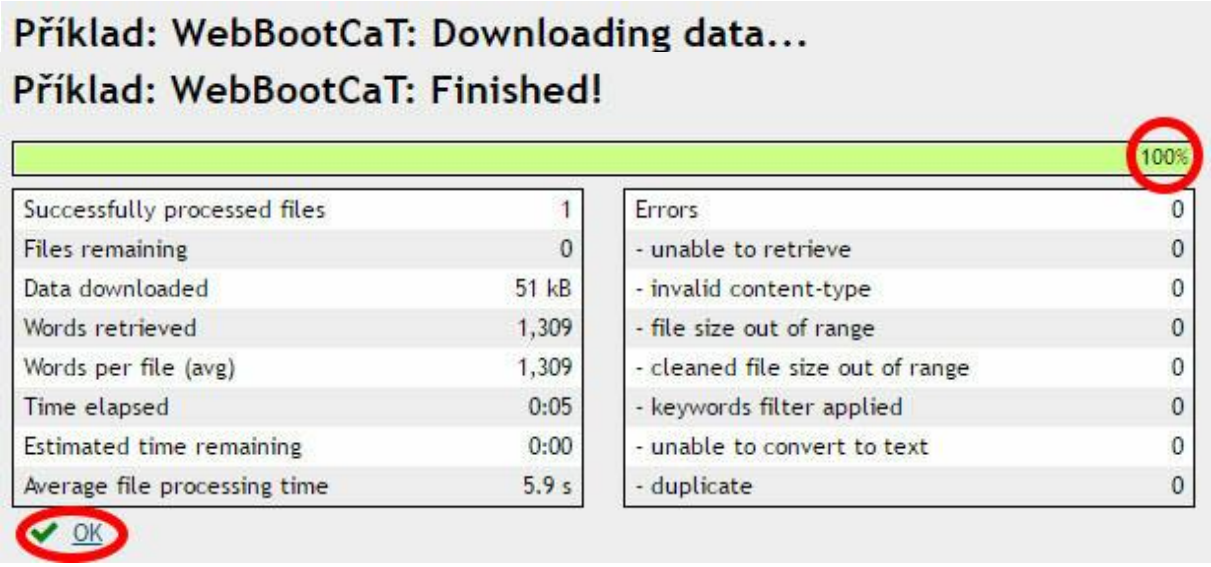
(unikátní alfanumerické ID korpusu bude vygenerováno automaticky)

2. jazyk korpusu (Sketch Engine bude s korpusem zacházet podle příslušných nastavení pro češtinu nebo jiný jazyk); pokud jazyk našeho korpusu není mezi zobrazenými možnostmi, zaškrtneme *other (UTF-8)*

3. Jeden ze způsobů definování korpusu:

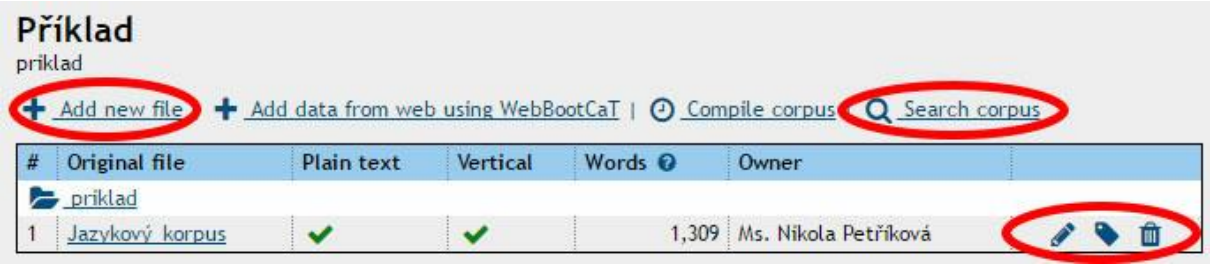
- Seed words: vepíšeme klíčová slova definující téma korpusu (která můžeme později změnit, přidat či odebrat)
- URLs: vepíšeme seznam URL adres, ze kterých se budou data pro vznik korpusu stahovat
- Web site: webová stránka, která se stáhne do korpusu kompletně celá, pokud obsahuje maximálně 2000 textových dokumentů

Po zadání všech potřebných informací klikneme na *Next* a Sketch Engine začne následně stahovat data ze stránek, které jsme uvedli. Tento proces trvá několik sekund až několik minut, a to v závislosti na množství stahovaných dat. Stahování je dokončeno ve chvíli, kdy indikátor v zeleném poli dojde do 100 % a zobrazí se oznámení *Finished!*.



Obrázek 26

Po kliknutí na *OK* vidíme uživatelské rozhraní s informacemi o korpusu. V tomto rozhraní můžeme v korpusu vyhledávat (Search corpus), spravovat jej, rozšiřovat nebo mazat (ikonky na pravé straně tabulky).



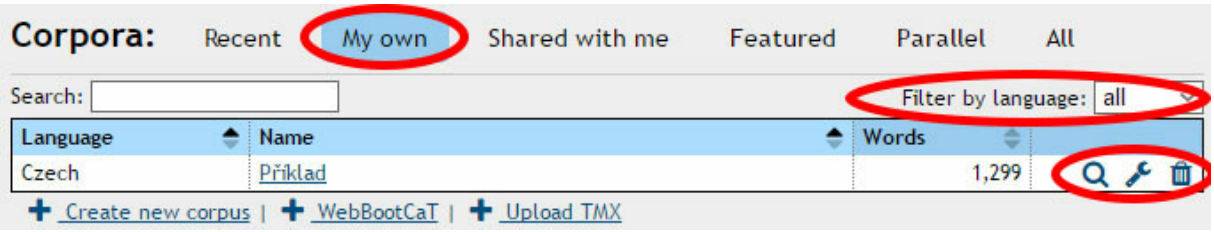
Obrázek 27

Pro zobrazení vlastního korpusu zvolíme na domovské stránce Sketch Engine záložku *My own*. Nyní můžeme tento korpus používat stejně jako kterýkoli jiný z nabídky Sketch Engine. Vlastní korpusy můžeme tímto způsobem také spravovat, třídit podle jazyka,

aktualizovat nebo mazat (ikonky na pravé straně tabulky).

1.10. Extrakce klíčových slov a termínů z korpusu

Tato funkce funguje pouze pro korpus, který vytvořil sám uživatel. Jak si vytvořit vlastní korpus popisujeme v kapitole Vytvoření vlastního korpusu (viz výše).



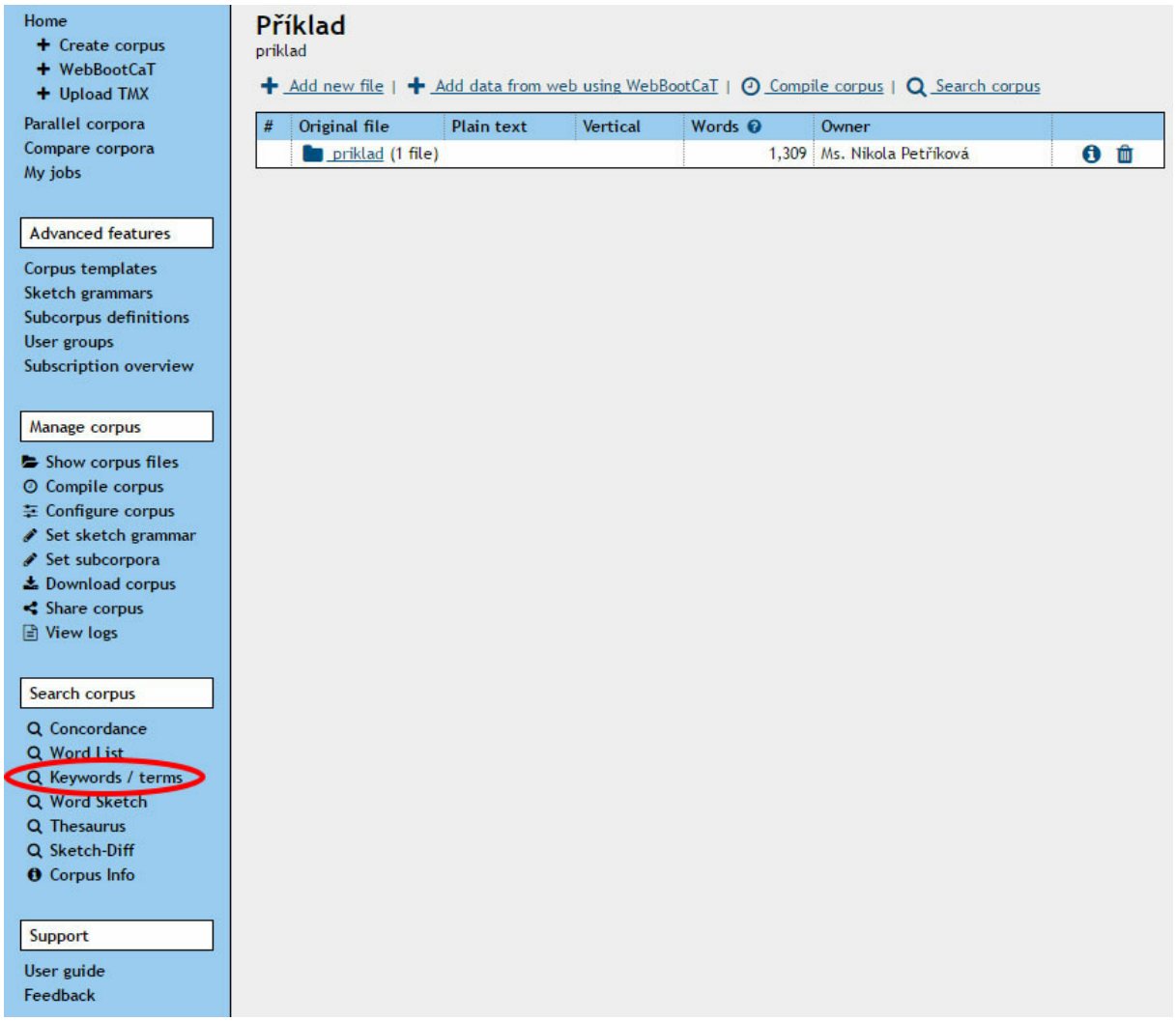
Obrázek 28

Sketch Engine má funkci extrakce klíčových slov a termínů, která pracuje pomocí přiřazování vzorců vyhledaných v textu a počítání frekvence výskytů určitých jazykových jevů. Tento algoritmus navíc bere v potaz jazyková data a informace o textu z hlediska lingvistiky (což je velmi cenný nástroj pro práci překladatelů či terminologů).

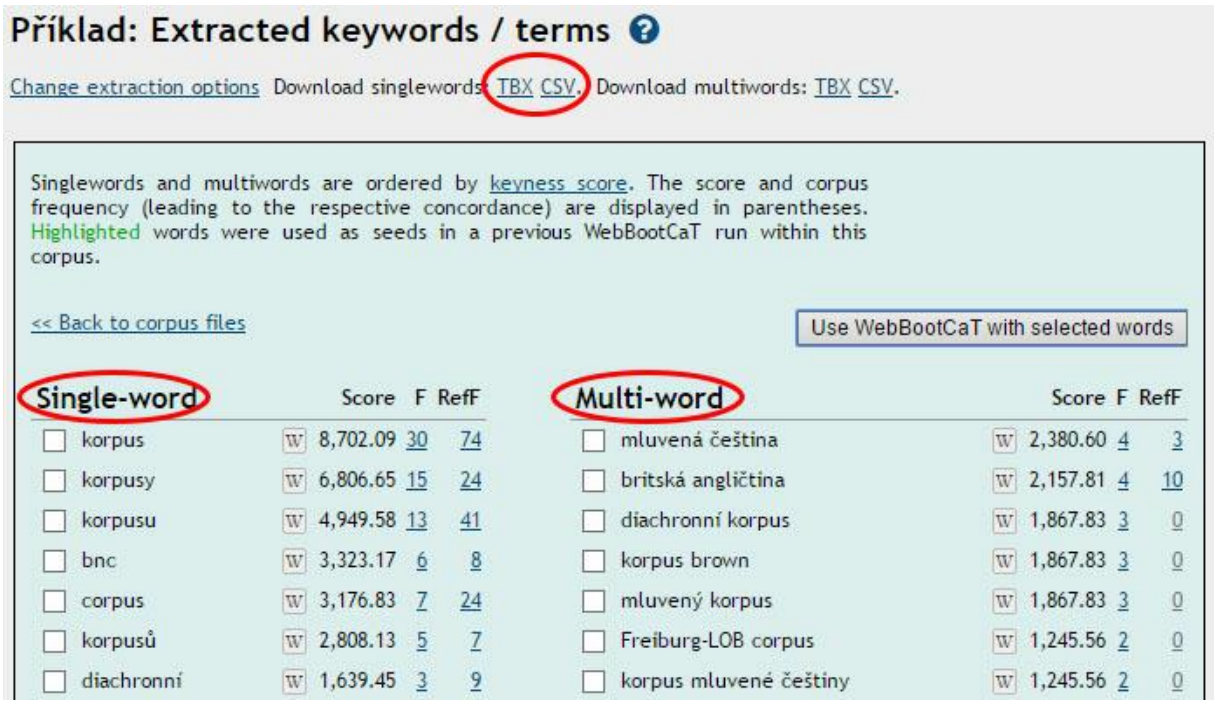
Klíčová slova a termíny z korpusu extrahujeme po kliknutí na ikonu klíče v tabulce zobrazující náš korpus, následkem čehož se zobrazí následující rozhraní.

Vybereme v menu položku *Keywords / terms*. Proces extrakce začne automaticky a trvá několik vteřin. Jeho výsledkem je výpis jednoslovných (první sloupeček) a víceslovných slovních spojení (druhý sloupeček), které se v korpusu objevují.

Každý termín má vedle sebe zaškrtávací políčko a uživatel si tak může sám vybrat, které pojmy budou jeho korpus reprezentovat. Modrým písmem na každém řádku je zobrazen počet užití konkrétní podoby slova nebo spojení v textu a taktéž obsahující odkaz na příklady v korpusu.



Obrázek 29



Obrázek 30

Výpis je možné stáhnout do počítače prostřednictvím odkazů v horní části obrazovky:

- Odkaz TBX slouží k importu výsledků do formátu CAT (tento software pomáhá překladatelům při pořizování překladů s udržováním konzistence v terminologii a také podporuje rychlost a efektivitu procesu překladu prostřednictvím návrhů na opětovné použití pasáží, které již autor překládal v minulosti. Je možné jej nastavit i tak, aby takto rozpoznané části textu automaticky nahradil dříve použitým překladem)
- Odkaz CSV pak umožňuje otevření ve formátu Microsoft Excel. Obě tyto funkce je možné použít pro stahování seznamů jak jednoslovných, tak víceslovných slovních spojení

Extrakce termínů je dostupná pro předem načtené korpusy přes menu *Word list*. Výsledky extrakce bilingválních slovních spojení z paralelních korpusů (například z těch, které jsou vytvořeny na základě překladů z nástroje CAT) mohou sloužit jako podklad pro tvorbu automaticky generovaných bilingválních glosářů.

1.11. Slovní seznamy

Sketch Engine je dále schopen vygenerovat následující druhy slovních seznamů:

- Seznam všech slov v korpusu
- Slova začínající na / končící na / obsahující určité znaky
- Seznam podstatných jmen / přídavných jmen /sloves nebo kterýchkoli jiných slovních druhů

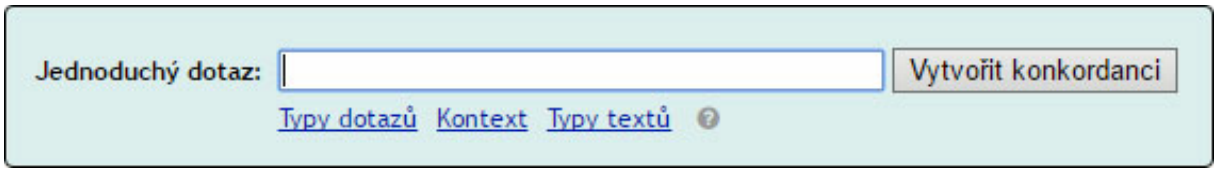
1.12. Konkordance

1.12.1. Hledání příkladů užití slov / slovních spojení v kontextu

Konkordance se používá při zkoumání kontextu konkrétního slova (lemmatu, fráze, …) a jejím výstupem je seznam všech výskytů tohoto slova (které také označujeme jako klíčové slovo nebo dotaz, anglicky keyword nebo query) v korpusu. Hledaný termín je zobrazen ve středu obrazovky a zleva i zprava je doplněn svým bezprostředním kontextem. Nejčastěji se setkáme s formátem konkordance, který se nazývá KWIC (zkratka pro Key Word In Context, česky: klíčové slovo v kontextu, viz slovník pojmů). Tento formát nám umožňuje prohlížet a porovnávat výskyty konkrétních slov, lemmat, slovních spojení nebo i komplexních slovních struktur v rámci svých jedinečných kontextů. Zkratkou KWIC také můžeme označit samotné slovo v centru konkordance, tedy výsledek našeho dotazu.

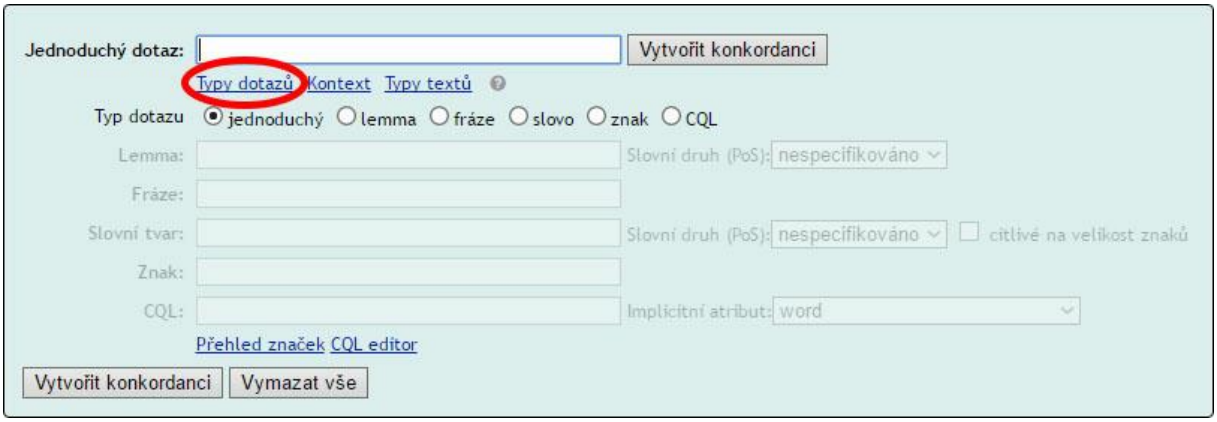
Po přihlášení si nejprve zvolíme vhodný korpus klinutím na jeho název.

Poté do prázdné kolonky vepíšeme požadovaný dotaz a klikneme na *Vytvořit konkordanci*.



Obrázek 31

Při zadávání dotazů pro vytváření konkordancí Sketch Engine nabízí více možností formulování dotazů, přičemž výsledky různých typů dotazů se mohou lišit. Kliknutím na *Typy*



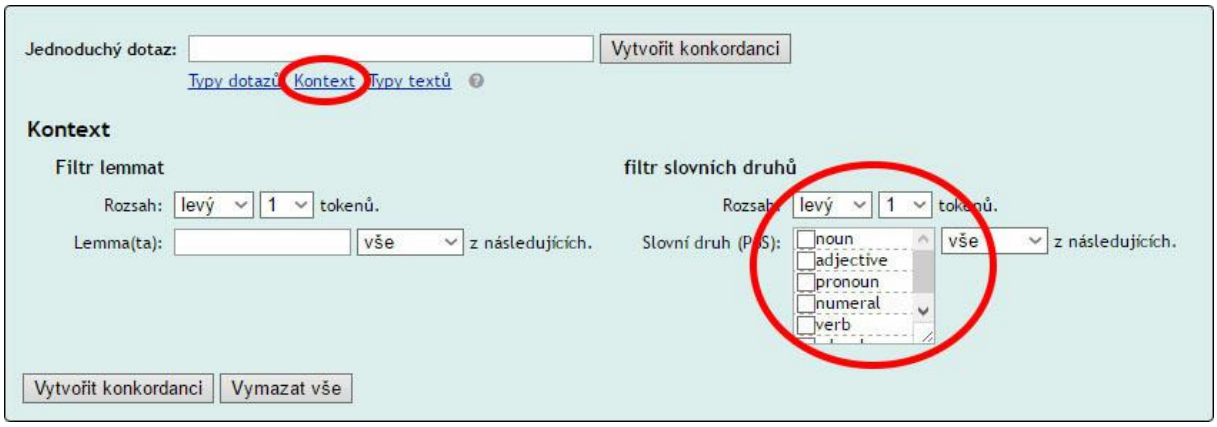
Obrázek 32

dotazů se zobrazí rozšířené možnosti vyhledávání a pak můžeme zvolit kteroukoli z následujících forem zápisu dotazu (pokud to vlastnosti korpusu, ve kterém konkordanci vyhledáváme, umožňují):

- Jednoduchý = vyhledává lemma, jedno i více slov, a to ve všech tvarech, ve kterých se vyskytují vedle sebe v korpusu. Je to nejméně specifikovaný druh dotazu, slouží pro základní potřeby vyhledávání konkordancí v korpusu a dá se nejlépe využít v případě, kdy nás zajímá, zda se námi zadaný dotaz vůbec v korpusu nachází a pokud ano, v jakém kontextu nejčastěji
- Lemma = základní forma slova (typicky užívaná například ve slovnících), která se používá pro vyhledávání všech tvarů a forem slova, které lze z tohoto základního tvaru generovat (rozlišuje použití velkých a malých písmen). Roletkou vedle pole, do kterého zadáváme dotaz, lze specifikovat slovní druh výsledků (např. při vyhledávání určitého významu homonymního slova)
- Fráze = vyhledávání dvou a více slov v korpusu, pouze v konkrétním tvaru, který jsme do dotazu zadali; bere ohled na použití velkých a malých písmen

- Slovo = vyhledává konkrétní tvar slova; je možné ovlivnit slovní druh výsledné konkordance a také, zda výsledek bude citlivý na velikost písmen nebo ne
- Znak = vyhledává znak nebo posloupnost znaků v korpusu, množství a opakování znaku nebo znaků zadaných v dotazu nehraje ve výsledné konkordanci roli
- CQL = Corpus Query Language je dotazovací jazyk používaný pro specifické vyhledávání v korpusech za použití regulárních výrazů. Využijeme jej hlavně v případě, že hledáme složité gramatické nebo lexikální vzory nebo potřebujeme použít taková kritéria vyhledávání, která nelze nastavit pomocí standardního uživatelského rozhraní

Pokud chceme zjistit, jaký kontext se nejčastěji pojí se slovem nebo slovním tvarem, který vyhledáváme, můžeme třídit výstup podle bližší specifikace předcházejícího a následujícího kontextu. Pod polem, do kterého vpisujeme dotaz, klikneme na volbu *Kontext* a zobrazí se nám rozšířená možnost zadání dotazu. Prostřednictvím této funkce je možné specifikovat, na jaký kontext se chceme u našeho dotazu zaměřit. Díky roletce v pravé dolní části vyhledávacího pole můžeme zaškrtnout, jaký konkrétní kontext nás zajímá a chceme jej vyhledat. Zvolíme stranu, ze které se kontext k dotazu připojuje, množství tokenů (token je nejmenší jednotka korpusu, nejčastěji každý slovní tvar nebo jednotka interpunkce, kromě mezery) a slovní druh kontextu.



Obrázek 33

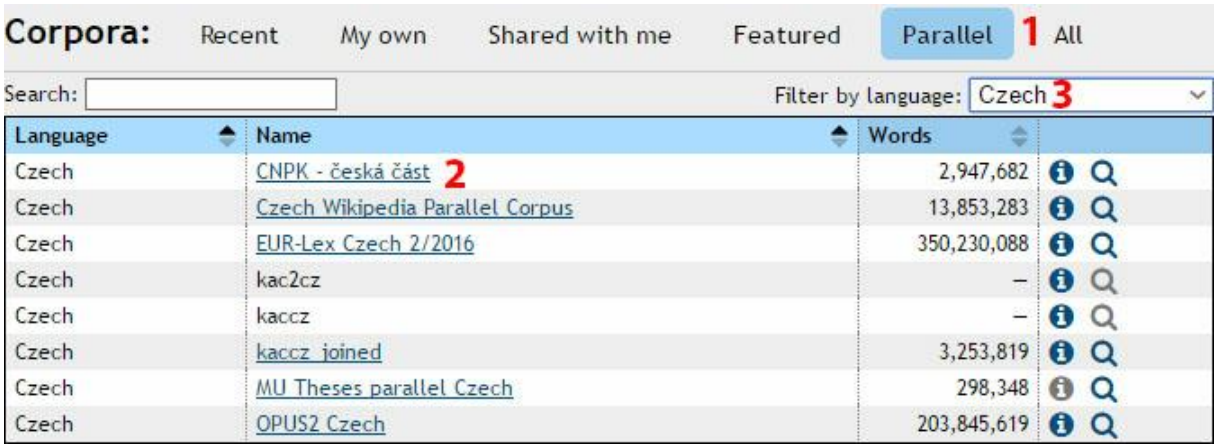
1.13. Paralelní korpusy

1.13.1. Práce s textem a jeho překladem

Mezi další vlastnosti Sketch Engine patří ty, které usnadňující práci překladatelům. Můžeme například zjistit, jak různí lidé překládají stejné texty z různých jazyků, nebo jak se překládají slova či slovní spojení, která v různých kontextech nesou různý význam. Prostřednictvím paralelních korpusů můžeme pracovat s dvojjazyčnými i vícejazyčnými texty,

a díky tomu vyhledávat a porovnávat slova či fráze a ukázky jejich překladů v kontextu.

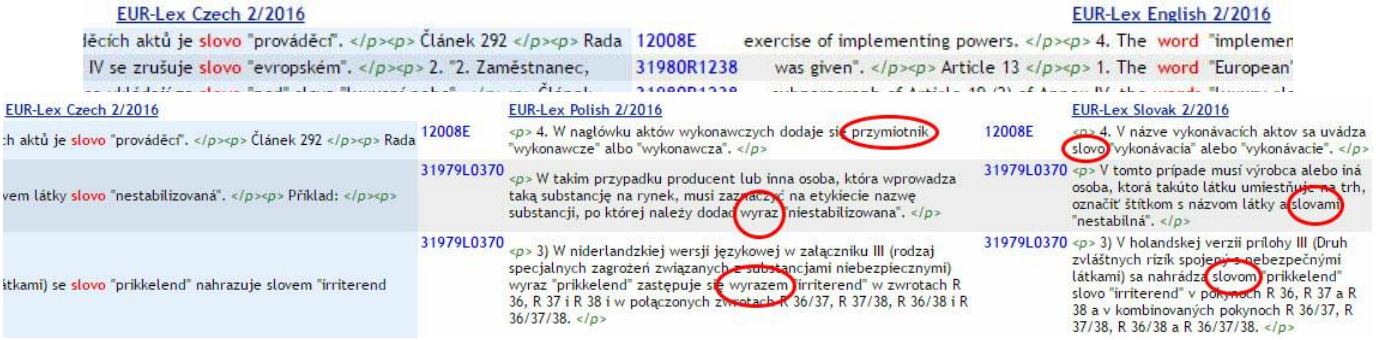
V rozhraní Sketch Engine jsou implementovány paralelní korpusy v mnoha jazycích, což poskytuje uživatelům zaměřujícím se na překlady široké pole působnosti. Taktéž je možné pracovat s vlastním vícejazyčným korpusem, vloženým např. z paměti nástroje CAT.



Obrázek 34

1. Po přihlášení si vytřídíme seznam korpusů na ty paralelní
2. Zvolíme korpus, který chceme použít kliknutím na jeho název
3. Výsledky vyhledávání je možné třídit podle jazyka korpusu

Vyhledávání funguje na principu seznamu konkordancí (viz kapitola Konkordance), můžeme tedy dostatečně specifikovat dotaz. Je možné zadat pojem v češtině i v dalším jazyce (v tomto případě angličtina), který si vybereme pro porovnání:



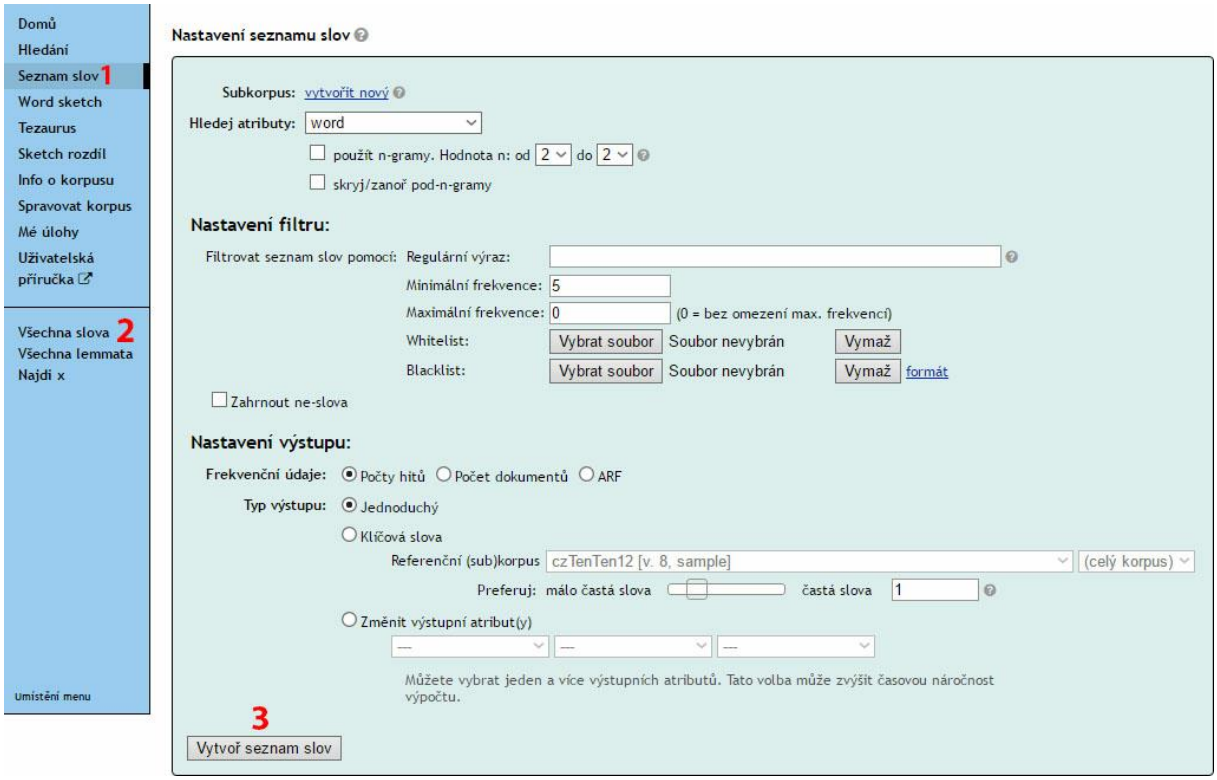
Obrázek 36

Nebo je možné zadat dotaz pouze v jediném jazyce a nechat vyhledávač jej přeložit (v tomto případě dotaz v češtině, výsledky v češtině, polštině a slovenštině).

Sketch Engine zobrazí výsledky vodorovně vedle sebe, nejčastěji na úrovni věty (to znamená, že navzájem sobě v překladu odpovídající věty budou zobrazeny vedle sebe). Hledané slovo je v mateřském jazyce zvýrazněno červeně, ale v dalších jazycích zvýrazněno není. Červeně vyznačené jsou přeložené výrazy pouze v případě, že do vyhledávače zadáme výraz jak v češtině, tak i ve druhém jazyce.

1.14. Seznam slov

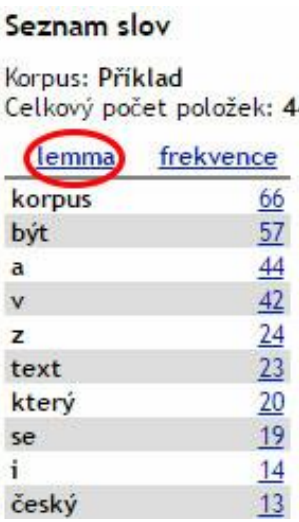
Tato funkce vygeneruje frekvenční seznam slov, který obsahuje všechna slova nebo lemmata nebo jakkoli jinak specifikované části textu, například podstatná jména, slovesa, přídavná jména, příslovce atd. Používá se pro zjištění množství konkrétních slov v korpusu.



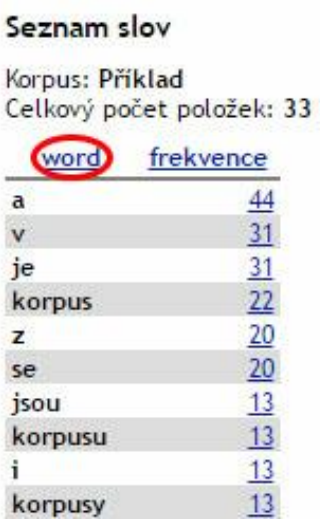
Obrázek 37

Pokud chceme tuto funkci využít, po přihlášení do manažeru Sketch Engine si vybereme korpus, ze kterého chceme seznam slov vygenerovat a klikneme na jeho název. Poté v levém menu klikneme na *Seznam slov* (1).

Pro základní účely postačí vědět, že v levém menu si můžeme vybrat z možností *Všechna slova* a *Všechna lemmata* (2) přičemž každá z těchto možností nám vygeneruje rozlišný výsledek (seznam slov nebo seznam lemmat). Volba *Všechna slova* bude mít na výstupu stejný seznam slov jako ten, který se zobrazí po kliknutí na *Vytvoř seznam slov* (3).



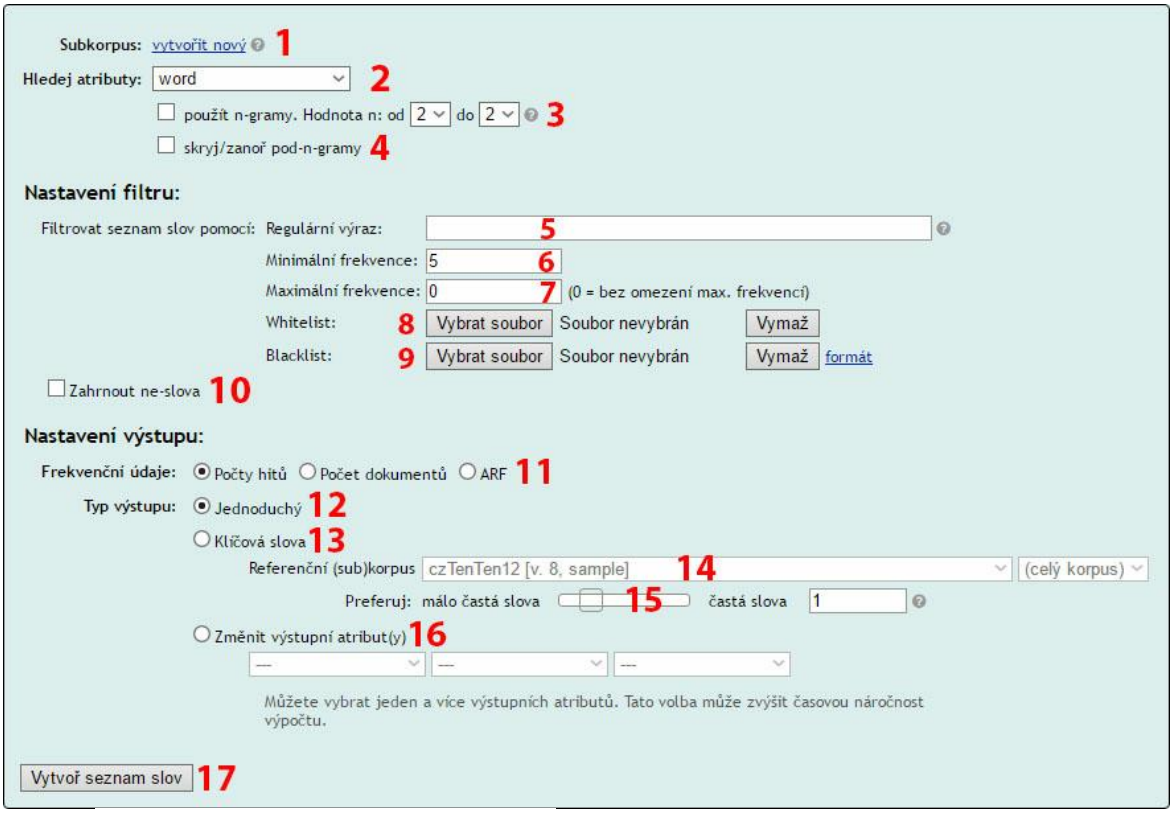
Obrázek 38



Obrázek 39

Funkce seznam slov ale může generovat i další druhy frekvenčních seznamů, např.:

- Seznam morfologických tagů, které se nejčastěji vyskytují v korpusu
- Seznam všech znakových trigramů v korpusu
- Seznam všech slov v korpusu, která začínají na *b*
- Seznam všech slovních variant slovesa *být*



Obrázek 40

Takovéto seznamy mohou být vygenerovány za použití formuláře, který se zobrazí po kliknutí na *Seznam slov*.

1. Seznam slov může být vygenerován z celého korpusu, nebo jen z jeho části, tzv. subkorpusu (pokud ten je k dispozici; v případě, že k dispozici není, můžeme odkazem *vytvořit nový*)
2. Touto funkcí vybíráme to, co chceme v korpusu spočítat (slova, lemmata nebo jiné atributy). Možnosti se odvíjí podle toho, jak je korpus anotován, ale nejčastěji zahrnují následující:
    - Atributy: tag, slovní tvar, lemma, lempos (lemma + part-of-speech, česky slovní druh)
    - Word Sketch: kolokace a pojmy – stejný princip jako u extrakce klíčových slov a termínů z korpusu, výsledky jednoslovných i víceslovných termínů se ovšem generují do jedné společné tabulky
    - Typy textu: závisí na zvoleném korpusu a typu textu v něm použitém

(typy textů dělí korpusy například podle zdroje na knižní, novinové atp. či podle média na mluvené a psané)

3. Tuto možnost zaškrtneme, pokud chceme spočítat frekvenci n-gramů
4. Když je tato možnost zaškrtnutá, budou sub n-gramy seřazeny pod n-gramy s vyšší hodnotou (např. 3-gram *at the end* bude zařazen pod 4-gram *at the end of*)

Nastavení filtru (vyloučí z výsledku všechny položky, o které nemáme zájem)

5. Omezení výsledku za pomoci regulárních výrazů (pro představu například regulární výraz *ka** vygeneruje seznam všech slov na ta, která začínají písmeny *ka*)
6. Limit pro vyloučení nízkofrekvenčních slov (tzn. slov, která se v korpusu vyskytují velmi málo, tudíž nemusí být zcela relevantní). Pokud vepíšeme nulu, ve výsledku se zobrazí všechna slova bez filtrování.
7. Limit pro vyloučení vysokofrekvenčních slov (tzn. slov, která se v korpusu vyskytují velmi často)

8. Whitelist použijeme, pokud chceme generovat výsledky pouze z určitého uzavřeného seznamu slov. Prostřednictvím této volby můžeme tento vlastní seznam do Sketch Engine nahrát, za předpokladu že se jedná o čistý UTF-8 formát souboru s jedním slovem na řádek a s položkami korespondujícími s vybranými atributy. Například pokud je za atribut zvoleno lemma, dotaz ve formě slova *šel* nevygeneruje žádný výsledek, jelikož se nejedná o lemma (*jít*) atp.
9. Blacklist použijeme, pokud chceme z vyhledávání vyloučit určitý uzavřený seznam slov
10. Pokud zaškrtneme tuto volbu, budou ve výsledném seznamu slov zahrnuta i takzvaná *ne-slova*, což jsou tokeny, které nezačínají na písmeno (například číslice nebo interpunkce)

Nastavení výstupu (specifikuje, co konkrétně se zobrazí na výstupní obrazovce po vygenerování seznamu slov)

11. Frekvenční údaje lze rozdělit na:
    - počty hitů (vedle každé vyhledané položky bude zobrazen počet jejích výskytů)
    - počet dokumentů (počet dokumentů v korpusu, ve kterých se vyhledaná položka vyskytla alespoň jednou)
    - ARF = Average Reduced Frequency, česky průměrná snížená frekvence (speciální statistika, která nepočítá vícenásobné výskyty téhož slovního tvaru, pokud se vyskytují příliš blízko u sebe, například v rámci jednoho dokumentu)
12. Jednoduchý typ výstupu (tato volba vytvoří seznam slov skládající se z těch položek, které odpovídají všem zadaným kritériím)
13. Klíčová slova (tato volba do výsledného seznamu slov zahrne pouze klíčová slova, například pouze specializovanou terminologii z určité vědní oblasti, kterou se korpus zabývá)
14. Referenční subkorpus musí být vždy zvolen (pokud si nejsme jistí, neměníme přednastavenou možnost)
15. Tento posuvník se vztahuje k nastavení referenčního subkorpusu (viz výše) a ovlivňuje, jak moc běžná (tzn. málo specializovaná) slova mají být ve výstupu zobrazena

16. Výsledky mají být spočítány pro určité atributy, ale vyobrazeny mohou být pro atributy odlišné, například můžeme vygenerovat frekvence konkrétních slovních tvarů, ale na výstupu si můžeme nechat zobrazit lemmata (můžeme nechat zobrazit až tři atributy)
17. Volba pro provedení výpočtu seznamu slov

1.15. FAQ

Přeložené některé užitečné Často kladené otázky z oficiálního anglického manuálu na webových stránkách Sketch Engine.

1.15.1. Zkušební účet (tzv. trial)

a) Je tento účet skutečně zdarma?

- Ano, je. Pokud se zaregistrujete jako uživatel trialu, nebudou po vás vyžadovány žádné platby ani data ke kreditní kartě.

b) Co všechno budu mít ve zkušební verzi Sketch Engine k dispozici?

- Budete mít přístup kompletně ke všem funkcím Sketch Engine (to znamená konkordance, Word Sketch, tezaurus, slovní seznamy atd.) a také ke 100 hotových korpusů ve více než 75 jazycích.

c) Můžu si ve zkušební verzi vytvořit svůj vlastní korpus?

- Ano, budete mít přístup ke Korpusovému architektu a budete mít možnost si vytvořit vlastní korpus prostřednictvím nástroje WebBootCaT.

d) Pokud se rozhodnu si po skončení zkušebního užívání pořídit placenou verzi Sketch Engine, zůstane můj vlastní korpus (vytvořený v trialu) zachován?

- Ano, převede se na účet, který si následně vytvoříte.

e) Můžu si vyzkoušet práci se Sketch Engine zcela bez registrace?

- Ano, ale vaše možnosti budou velice omezené. Budete mít přístup pouze ke třem korpusům v anglickém jazyce.

1.15.2. Druhy předplatného

a) Není mi jasné, který typ předplatného je pro mne nejvhodnější.

- Akademické předplatné je vhodné pouze pro potřeby osob a institucí, které fungují nekomerčně a ani nejsou spřátelené s nějakou výdělečnou společností.

- Neakademické předplatné je vhodné pro osoby, instituce i společnosti, které budou Sketch Engine využívat pro komerční účely.
- Omezené komerční předplatné mohou využívat překladatelé na volné noze, terminologisté nebo copywriteři, kteří nemají dlouhodobý závazek vůči výdělečné společnosti.
- Plné komerční předplatné je pak vhodné pro lexikografy a všechny výdělečné organizace.

b) Můžu změnit frekvenci, s jakou platím své předplatné?

- Ano, můžete měnit mezi tříměsíčním a ročním placením předplatného. Změna se projeví po vypršení vašeho současného typu předplatného. Pro změnu způsobu placení účtu nás prosím kontaktujte zde: https://www.sketchengine.co.uk/request-support/

c) Co znamená předplatné typu učitel + žák?

- Toto předplatné poskytuje učitelům přístup do Sketch Engine pro potřeby svých žáků, a to za zvýhodněnou cenu. Podmínky používání platí. Více v ceníku: https://www.sketchengine.co.uk/price-list/

1.15.3. Uživatelský prostor, předplatné, platby a faktury

a) Jak zjistím stav svých faktur a předplatného?

- Podívejte se do záložky Subscription Overview na stránkách Sketch Engine.

b) Kde najdu fakturu své platby?

- Podívejte se do záložky Subscription Overview na stránkách Sketch Engine. Pokud ji tam nenajdete, kontaktujte nás na e-mailu inquiries@sketchengine.co.uk.

c) Jak můžu navýšit svůj uživatelský prostor?

- Předplatné hrazené kreditní kartou: podívejte se do záložky Subscription Overview na stránkách Sketch Engine. Zkontrolujte, že volba automatického obnovení je nastavená jako aktivní. Poté pro nákup většího množství prostoru pro Vaše korpusy, klikněte na *More space*.
- Předplatné hrazené bankovním převodem: kontaktujte nás prosím na e-mailu inquiries@sketchengine.co.uk.

d) Potřebuji změnit fakturační adresu.

- Podívejte se do záložky Subscription Overview na stránkách Sketch Engine a tam zvolte možnost Změnit fakturační adresu a nastavit vše potřebné. Pokud nastanou jakékoli problémy, prosíme, kontaktujte nás na e-mailu inquiries@sketchengine.co.uk.

1.15.4. Korpus a jazyky

a) Jak velké jsou korpusy ve Sketch Engine?

- Velikost korpusů se liší kus od kusu, ten největší momentálně obsahuje až 20 miliard slov.

b) Které jazyky podporujete?

- Sketch Engine momentálně podporuje více než 80 jazyků, jejich kompletní seznam najdete zde: https://www.sketchengine.co.uk/user-guide/user-manual/corpora/by-language/

c) Můžu sem nahrát vlastní dokumenty a vytvořit z nich korpus?

- Ano, můžete. Podrobný popis postupu, jak na to se nachází zde: https://www.sketchengine.co.uk/user-guide/user-manual/corpora/create-from-files/

d) Plánujete rozšířit množství použitých jazyků?

- Ano, neustále se snažíme Sketch Engine obohacovat o nové korpusy a nové jazyky. Pokud se chcete jako první dozvědět o novinkách probíhajících při rozšiřování manažeru, sledujte nás na Facebooku, Twitteru nebo se zaregistrujte k odběru našich novinek.

e) Můžu sem nahrát dokumenty v jazyku, který Sketch Engine momentálně nepodporuje?

- Ano, budete mít možnost vyhledávat v korpusu, vygenerovat seznamy slov i kolokace ale některé z pokročilých možností nebudou k dispozici.

1.15.5. Technické dotazy

a) Musím si něco nainstalovat?

- Nemusíte, Sketch Engine funguje online a je podporován většinou standardních internetových prohlížečů.

b) Jaké jsou minimální technické požadavky?

- Žádné. Bohatě stačí připojení k internetu a webový prohlížeč.

c) Můžu si stáhnout vlastní Sketch Engine?

- Tuto možnost nabízíme větším institucím nebo redakcím, a to v případě, že není možné jejich korpusy z legálních důvodů umístit na naše servery. Mějte na paměti, že taková instalace obnáší přenosy terabytů dat.

2. Slovník pojmů

Přeložený glosář, který je součástí oficiálního anglického manuálu na webových stránkách Sketch Engine.

ARF neboli Average Reduced Frequency (průměrná redukovaná frekvence)

Speciálně upravená funkce, která zabraňuje tomu, aby byl výsledek vyhledávání ovlivněn výsledky z pouze jedné části korpusu (například pouze z jednoho dokumentu). Pokud se vyhledávaný výraz v určité části korpusu vyskytuje hojněji než v jiných částech, mohlo by to ovlivnit výsledky vyhledávání. ARF tomuto ovlivnění zabraňuje. Pokud je výraz v korpusu zastoupen rovnoměrně, je frekvence ARF srovnatelná s Frekvencí v milionu.

CAT tool (nástroj CAT)

Počítačem řízený pomocný software. Nástroj, který pomáhá překladatelům sjednotit terminologii v jejich překladatelských pracích a také je nápomocný při samotném překladatelském procesu. Funguje tak, že navrhuje (nebo automaticky překládá či nahrazuje přeložená slova) způsob, jakým přeložit určité slovo nebo část výpovědi, kterou překladatel již v minulosti překládal. Udržuje tím jednotnost a srozumitelnost odborného textu.

Cluster

Jako cluster označujeme proces vytváření skupin slov v tezauru nebo slovních profilech (Word Sketch). Slova jsou k sobě navzájem přiřazena podle společných rysů jejich chování v rámci kolokace.

Collocate (kolokát)

Termínem kolokát se označuje ta část kolokace, která není uzlem. Termínem uzel označujeme hlavní část kolokace, ke které se přidružují právě tyto kolokáty. Například v kololaci *silný vítr* je kolokát *silný* a uzel *vítr*.

Collocation (kolokace)

Kolokací (dříve se užíval termín slovní spojení) se v rámci korpusové lingvistiky nejčastěji rozumí smysluplné, ustálené, syntagmatické spojení dvou (nebo víc) slovních tvarů (někdy celých lexémů) v blízkém kontextu (ČNK: Kolokace, 2016). Například kolokace *fatální chyba* se nejčastěji skládá z uzlu (*chyba*) a kolokátu (*fatální*). Kolokace se různí svou silou, například slovní spojení *hodně krve* je velmi slabé, protože oba výrazy se velmi často pojí s jinými slovy, zatímco *tratoliště krve* je velmi silné spojení, protože se tato slova velmi často vyskytují v kontextu současně.

Concordance (konkordance)

Konkordance je seznam všech příkladů užití určitého slova nebo fráze, které se nám podařilo

vyhledat v korpusu. Nejčastěji se jedná o konkordance typu KWIC (viz níže heslo KWIC), ve kterých vidíme vyhledávané slovo zvýrazněné uprostřed obrazovky, obklopené svým kontextem zleva i zprava.

Concordancer

Program, který zobrazuje konkordance, což jsou výsledky vyhledávání slovních spojení v korpusech prostřednictvím korpusového manažeru Sketch Engine.

Corpus (korpus)

Objemný soubor obsahující velké množství textu pro diachronní studium jazyka. Vlastností korpusu je to, že je označkovaný (tzn. je opatřen speciálními značkami neboli tagy, které jednoznačně určují slovní druh a gramatické kategorie všech slov v něm obsažených). Termíny korpus, textový korpus a jazykový korpus jsou mezi sebou zaměnitelné. Jazykové korpusy jsou neocenitelným přínosem pro zkoumání jazyka v jeho reálné, přirozené a každodenní podobě.

Corpus architect (korpusový architekt)

Jedna s intuitivní funkcí Sketch Engine. Nástroj, který se používá pro vytváření korpusů z dokumentů stažených z webu, které nevyžadují žádné speciální znalosti problematiky textu.

Corpus manager (korpusový manažer)

Korpusový manažer je počítačový program používaný k práci s korpusy, například k jejich vytváření, kompozici, editaci, anotování a vyhledávání v nich. Sketch Engine je uživatelské rozhraní určené k práci s korpusovým manažerem Manatee.

CQL neboli Corpus Query Language (korpusový dotazovací jazyk)

CQL je systém zadávání složitých dotazů pro vyhledávání komplexnějších výsledků v korpusu. Používá se v situacích, kdy není možné dotaz zadat prostřednictvím standardních uživatelských kritérií. Tato kritéria mohou zahrnovat i jiné atributy než slova, lemmata, tagy nebo typy textu. Je možné použít například logické operátory AND, OR nebo NOT.

CSV neboli Comma-separated values (hodnoty oddělené čárkami)

CSV je zkratka označující druh prázdného textového dokumentu, který se používá pro ukládání tabulkových dat. Tento formát je ideální pro export výsledků získaných prostřednictvím Sketch Engine, jelikož jej akceptují nejrůznější další aplikace na zpracování nebo zobrazení dat (např. Microsoft Excel, Open Office, Google Documents a mnohé další).

Disambiguation (desambiguace)

Proces identifikace jednoznačného významu slova, které může mít v jazyce více významů. Výsledkem desambiguace je konkrétní slovo svázané se svým jednoznačným významem.

Distributional thesaurus (distribuční tezaurus)

Automaticky vytvářený tezaurus, který vyhledává a řadí do skupin slova, která mají tendenci vyskytovat se v podobných kontextech jako námi zadaný dotaz. Není to totéž jako člověkem vytvořený tezaurus synonym, tento tezaurus se vytváří automaticky na základě algoritmů, které v korpusu vyhledávají slova vyskytující se v podobném kontextu.

Frequency per million neboli freq/mill (frekvence v milionu)

Počet výskytů dotazu v milionovém korpusu. Používá se pro porovnávání frekvencí výskytů slov v korpusech rozdílných velikostí. Vzorec pro spočítání frekvence v milionu:
počet výskytů / velikost korpusu v milionech = frekvence v milionu
Například token, který jsme našli desetkrát v milionovém korpusu, bude mít frekvenci v milionu 10. Token, který jsme našli stokrát ve sto milionovém korpusu, bude mít frekvenci v milionu 1. To znamená, že druhý zmíněný je méně často se vyskytující token.

GDEX neboli Good Dictionary Examples (dobré příklady ze slovníku)

Věty, které mohou být užitečné jako příkladové věty ve slovníku (například proto, že jsou dostatečně názorné a reprezentativní). Výsledek nástroje GDEX můžeme zobrazit po vyhledávání dotazu v korpusu kliknutím na položku *Možnosti zobrazení* v levém menu a následně zaškrtnutím *Třiď dobré slovníkové příklady* (GDEX). Množství výsledků lze ovlivnit číslem vepsaným do pole na konci následujícího řádku: *Zobraz GDEX skóre v konkordanci (pomalé, používat pouze pro vývoj GDEX) Počet řádků k seřazení: 100*.

Global subcorpus (globální subkorpus)

Globální subkorpus je takový druh subkorpusu, který je možné sdílet se všemi uživateli Sketch Engine bez rozdílu. Takto sdílet subkorpus je možné za pomoci nástroje Korpusový architekt nebo prostřednictvím skriptu mksubc.py.

Header field (záhlaví)

Obsahuje nejrůznější informace vztahující se k dokumentům v korpusu. Například korpus, který obsahuje dokumenty stažené z internetu z různých domén, může být řazen podle těchto domén za použití záhlaví <doména dokumentu> a hodnoty „názevdomény“ (<doména dokumentu>=„názevdomény“).

KWIC neboli Key Word In Context (klíčové slovo v kontextu)

KWIC je nejčastěji používaný formát zobrazení konkordancí vyhledávaných v korpusu. V tomto formátu se vyhledávaný termín nebo fráze zobrazí uprostřed obrazovky (obvykle zvýrazněn) a je obklopen svým kontextem zprava i zleva. Termín KWIC je často užíván také pro označení zvýrazněného slova v centru konkordance, tedy jako výsledek vyhledávání.

Lc neboli Lower case (malé písmeno)

Zkratka lc označuje slovo začínající malým písmenem. Používá se pro jednoznačné rozlišení vyhledávaného termínu v rozhraních zohledňujících velká a malá písmena (např. *Hrad* není totéž, co *hrad*).

Learner corpus

Tento korpus vzniká na základě dat poskytnutých osobami, které se daný jazyk teprve učí. Jeho účelem je interpretovat nejčastější chyby a omyly nastávající při studiu jazyka. Takový typ korpusu je možné ve Sketch Engine použít jako zdroj omylů i jako zdroj oprav anotací. Speciální vyhledávací rozhraní umožňuje vyhledávat podle předchozí verze (před opravou), následující verze (po opravě) nebo obou verzí.

Lemma

Lemma je základní podoba slova v jazyce. Lemmata nejčastěji nalezneme ve slovnících a vyhledávání podle lemmat ve výsledku zahrne jak základní podobu slova, tak všechny jeho další tvary. Například při vyhledávání lemmatu *pes* se ve výsledcích zobrazí *pes, psa, psu, psovi, psem* atd. Vyhledávání podle lemmat je citlivé na použití velkých a malých písmen, což znamená, že vyhledávání slova *pes* a *Pes* poskytne rozlišné výsledky.

Lemma_lc

Lemma_lc je druh lemmatu, který stírá rozdíly ve velikosti počátečního písmene. Všechna lemmata s velkým počátečním písmenem budou ve výsledcích vyhledávání převedena na malé, tedy Pes i pes budou totožné.

Lemmatization (lematizace)

Proces přiřazení lemmatu každému slovnímu tvaru v korpusu za použití automatizovaného nástroje, kterému se říká lemmatizér. Výsledek lemmatizace (tedy korpus obsahující základní tvary všech tokenů) je vhodné použít v situacích, kdy potřebujeme v korpusu vyhledat všechny tvary zadaného slova zároveň. Například při vyhledávání slova *pes* se ve výsledcích objeví jak základní tvar (lemma) *pes*, tak všechny jeho další tvary (*psa, psu, psovi, psem* atd.)

Lempos

Termín označující spojení lemmatu (*lem-*) a slovního druhu (anglicky Part Of Speech, tedy *pos*). Skládá se z pomlčky oddělující lemma a slovní druh, například *jít-v* (verb = verbum), dům-n (noun = substantivum). Zkratky pro označení slovního druhu se liší korpus od korpusu. Lempos je citlivý na velikost počátečního písmene, což znamená že *Hrad-n* a *hrad-n* není totéž.

lempos_lc

Druh lempos, který stírá rozdíly ve velikosti počátečního písmene. Všechny lempos s velkým počátečním písmenem budou ve výsledcích vyhledávání převedeny na malé, tedy *Hrad-n* a *hrad-n* budou totožné.

Likelihood (pravděpodobnost)

Porovnává frekvenci výskytu pojmu ve dvou korpusech. Kritérium mezní hodnoty by mělo být stejné pro oba korpusy. Ověřit, zda se pojem vyskytuje častěji v jednom korpusu než ve druhém, můžeme pomocí statistického testu log-likelihood.

Log-likelihood (logaritmus pravděpodobnosti)

Používá se k vypočítání pravděpodobnosti výskytu určitého slova nebo pojmu v jednom korpusu v poměru ke druhému.

logDice

Statistika závislá na míře frekvence slov w^1 a w^2 a bigramu w^1 w^2 v korpusu. Výpočet není ovlivněn velikostí korpusu.

Longest-commonest match (nejdelší nejčastější shoda)

Koncept zavedený Adamem Kilgarriffem a využívaný k označení těch nejčastějších párů kolokací, tedy těch částí jazyka, ve kterých se kolokace vyskytují nejčastěji. Nejdelší nejčastější shoda se vyskytuje na obrazovce výsledků vyhledávání ve Slovních profilech (Word Sketch), kde interpretuje typické chování vyhledaných kolokací.

Metadata

Strukturované informace o textech obsažených v korpusu. Metadatem je například lemma nebo název každého jednoho dokumentu obsaženého v korpusu.

Multilevel list (víceúrovňový seznam)

Seznam seřazený podle více než jednoho kritéria. Například frekvenční seznam seřazený podle slovních tvarů následovaných lemmaty a tagy.

N-gram

Pojem n-gram označuje řadu více za sebou následujících libovolných struktur (bigram = 2 struktury, trigram = 3 struktury, …, n-gram = n struktur). Typicky se jedná o písmena nebo slova, ale také může jít o fonémy nebo slabiky. Vytváření frekvenčního listu takových sekvencí napomáhá zjistit, jaké struktury v jazyce mají tendenci se na sebe vázat.
N-gramy se nejčastěji používají při vyhledávání Seznamů slov (word list).

Node (uzel)

V kontextu kolokací je uzel slovo, nacházející se uprostřed kolokace, například ve slovním spojení *silný vítr* je uzlem slovo *vítr* a slovo *silný* je kolokát.

V kontextu konkordancí uzlem chápeme slovo nebo slovní spojení, které také označujeme jako KWIC nebo jednoduše dotaz. Objevuje se v centru nebo zvýrazněné ve výsledcích vyhledávání konkordancí.

Non-word (ne-slovo)

Obecné označení pro všechny tokeny, které nezačínají písmenem z abecedy. Příkladem ne-slova může být například: *!important* nebo *2U*.

Overall score (celkové skóre)

Skóre jednotlivých vztahů, které závisí na výsledcích statistiky logDice ve Slovních profilech (Word Sketch). Toto skóre je zobrazeno v záhlaví každého sloupce zobrazujícího určitý vztah.

Parallel corpus (paralelní korpus)

Paralelní korpus je takový korpus, který se skládá ze dvou totožných textů zapsaných ve dvou (nebo více) různých jazycích. Texty jsou shodně zarovnané, často jsou stejné věty navzájem propojené odkazy. V korpusu se dá vyhledávat v jednom či obou jazycích pro snadné porovnání překladu.

POS neboli Part of Speech (slovní druh)

Zkratkou POS (nebo někdy PoS) označujeme slovní druhy, jakými jsou například substantiva, adjektiva, verba, adverbia a další.

POS tagger (značkovač slovních druhů)

Automatizovaný nástroj sloužící ke značkování slovních druhů všech tokenů v korpusu. Tag neboli značka nese informaci o slovním druhu tokenu a často také morfologickou a gramatickou informaci, jakou je například číslo, rod, pád, čas apod.

Positional attribute (poziční atribut)

Informace přidaná ke každému tokenu v korpusu, například jeho lemma (základní tvar slova) nebo slovní druh. Například pro slovo *psi* je lemma *pes*, tag *n* a lempos *pes-n*.

Atributy se mohou lišit korpus od korpusu, a to dokonce i mezi vícejazyčnými verzemi stejného korpusu.

Preloaded corpus (přednastavený korpus)

Korpus, který je hotový a připravený k použití, přestože nebyl vytvořen uživatelem (je tedy automaticky vygenerovaný), jedná se například o Britský národní korpus. Ve Sketch Engine jsou tyto korpusy dostupné předplatitelům a uživatelům trialu (zkušební verze).

Query (dotaz)

Posloupnost znaků či slov (nebo kombinace obojího), které zadáváme do Sketch Engine jako vstup, abychom na výstupu získali konkordanci. Při vyhledávání konkordancí nejčastěji

používáme jako dotaz slovní tvar, ale je možné vyhledávat i podle jiných parametrů a očekávat rozdílný výsledek. Sketch Engine nabízí mnoho druhů výstupů, například Slovní profily (Word Sketch), tezaurus, seznam slov apod.

Reference (odkaz)

Vlastnost dokumentu, která daný dokument reprezentuje. Například se jedná o URL adresu dokumentu. Každý dokument v korpusu obsahuje odkazy, které jej charakterizují.

Reference corpus (referenční korpus)

Korpus, který byl zvolen za standard pro porovnávání s jiným korpusem (zvoleným uživatelem). Tento korpus se používá pro klíčová slova.

Regular expressions (regulární výrazy)

Speciální symboly a znaky, které se používají při pokročilém vyhledávání v korpusech. Například pokud potřebujeme vyhledat všechna slova začínající na (nebo obsahující či končící na) určitý znak nebo posloupnost znaků, použijeme regulární výraz. Například při použití řetězce **ice* se na výstupu vyhledávání zobrazí všechna slova v korpusu končící na *-ice* která obsahují libovolný počet znaků.

Relative text type frequency (relativní frekvence typu textu)

Tato funkce porovnává frekvenci v určitém typu textu (část korpusu) s celým korpusem nebo porovnává frekvence v různých typech textů mezi sebou, a to i když nejsou stejně velké. Uživatel tak může pozorovat, zdali jsou vyhledávaná slova standardně používána v určitém typu textu. Hodnotu relativní frekvence typu textu spočítáme tak, že relativní frekvenci výsledků dotazů vydělíme relativní velikostí určitého typu textu. Výsledek můžeme chápat jako „jak moc/málo častý je výsledek dotazu v této části textu v porovnání s celým zbytkem korpusu“. Vyšší frekvence znamená vyšší hodnotu, větší část textu znamená nižší hodnotu. Například slovo test se v korpusu vyskytuje 2000krát. 400 z těchto výskytů je v „Mluveném“ typu textu a tento typ textu reprezentuje 10 % korpusu. Relativní frekvence typu textu je tedy 400 / 2000 / 0.1 = 200 %, a to znamená, že slovo test se dvakrát častěji vyskytuje v „Mluveném“ korpusu než v celém korpusu.

Salience (význam)

Statistická hodnota, která vykresluje význam určitého tokenu v rámci jeho konkrétního kontextu. Tato hodnota se počítá za pomoci logaritmu logDice.

Search attribute (vyhledávací atribut)

Dotaz, který zadáváme do Sketch Engine při vyhledávání a vytváření slovních seznamů (word list). Slovní seznam lze vytvořit pro slova, slovní tvary, lemmata, tagy a další.

Search span (rozsah vyhledávání)

Tento termín označuje počet tokenů na každé straně uzlu, které se budou navzájem shodovat pro potřeby filtrování výstupů konkordancí. Pokud hodnotu hledání nastavíme na −5 až 5, filtrujeme konkordance pouze na ty, které obsahují vyhledávaný výraz ve vzdálenosti jeden až pět tokenů před nebo za uzlem.

Simple math (jednoduchá matematika)

Jednoduchá matematika použitá ve Sketch Engine poskytuje jednoduché vzorce pro výpočet a identifikaci termínů a klíčových slov.

Structure (struktura)

Vlastnost korpusu, která označuje jeho rozdělení na části. Nejčastěji korpusy dělíme na části podle vět, odstavců nebo dokumentů, ale můžeme jej strukturovat také podle jiných parametrů závislých na typu a možnostech konkrétního korpusu.

Subcorpus (subkorpus)

Každý korpus lze rozdělit na libovolné množství menších částí. Těmto částem říkáme subkorpusy. Korpus můžeme na subkorpusy rozdělovat podle typů textu (například podle zdroje na knihy a časopisy, podle druhu na mluvený a psaný atd.) nebo podle jakýchkoli jiných uživatelem zvolených kritérií. Subkorpus můžeme taktéž vytvořit z konkordance, a tedy na základě výsledků vyhledávání v korpusu.

Tag (značka)

Tag neboli značka (také označovaná jako morfologická značka nebo POS tag) je označení přiřazené každému tokenu v anotovaném korpusu. Reprezentuje vlastnosti tokenu, jakými jsou například slovní druh nebo gramatické kategorie. Nástroj, který používáme pro značkování korpusu, se nazývá tagger a množinu všech tagů použitých v konkrétním korpusu označujeme slovem tagset.

Tagset

Tagset (nebo také tag set) je soubor všech značek použitých v jednom označkovaném korpusu.

Tick Box Lexicography

Zkratkou TBL označujeme aplikaci, kterou Sketch Engine používá pro shromažďování vět, jež se typicky uvádějí jako příklady v jazyce a mohou tedy sloužit jako podklad při tvorbě slovníků.

Term (termín)

Termínem může být klíčové slovo nebo více slov, které je v korpusu velice frekventované, a přitom se nejedná o běžně používané slovo (jakým je třeba *a*, *nebo* atd.). Tímto označením popisujeme výraz, který je charakteristický pouze pro daný korpus.

Term base (databáze termínů)

Ve spojení s nástrojem CAT pojmem databáze termínů rozumíme databázi specializované terminologie a dalších lexikálních jednotek, které je třeba překládat do jiného jazyka jednotně. Nástroj CAT používá databázi termínů pro kontrolu jednotnosti překladu, vyhledávání nepřeložených částí textu a návrh (nebo automatické nahrazení) překladů, které již byly použity v minulosti a jsou tedy zaneseny v této databázi.

Term extraction (extrakce termínů)

Proces identifikace jedinečných termínů, které jsou charakteristické pro určitý text a tím pádem jej specifikují. Používá se při tom specializovaný software, který vyhledává jednoslovné i víceslovné termíny které filtruje pro potřeby extrakce pomocí porovnávání více korpusů za použití referenčních korpusů (tedy extrahuje termíny, které se v jednom textu vyskytují hojně, ale nejedná se o běžné výrazy).

Text type (typ textu)

Označení konkrétního druhu textu vyskytujícího se v korpusu. Typ textu může odkazovat ke zdroji textu (noviny, kniha atd.), druhu korpusu (mluvený, psaný) nebo třeba k času či místu vzniku. Ne všechny korpusy mají dokumenty v nich obsažené opatřené označením typu textu. Korpus můžeme rozdělit na sub korpusy podle typů textu nebo podle nich můžeme generovat slovní seznamy.

Token

Nejmenší jednotka, na kterou můžeme korpus rozdělit. Nejčastěji se jedná o slovní tvary nebo interpunkci (čárky, tečky atd.) což znamená, že většinou korpus obsahuje více tokenů než slov. Mezery mezi slovy tokeny nejsou. Text korpusu můžeme rozdělit do tokenů prostřednictvím nástroje tokenizer, který je většinou přizpůsoben různým jazykům pro co nejrelevantnější výsledky (například don’t v angličtině se skládá ze dvou tokenů).

Tokenization (tokenizace)

Automatický proces rozdělování textu korpusu na jeho základní jednotky, tedy tokeny.

Tokenizer

Software, který se používá pro rozdělování textu korpusu na tokeny. Tokenizer je uzpůsoben jazyku korpusu, který tokenizuje, a zohledňuje tedy i všechny zvláštnosti jazyků. Například v angličtině je termín don’t tokenizován jako dva tokeny. Sketch Engine obsahuje tokenizery

mnoha jazyků a také jeden univerzální tokenizer, který se používá při tokenizaci korpusů v jazycích, které zatím nejsou ve Sketch Engine podporovány. Univerzální tokenizer využívá mezery mezi slovy pro odlišení jednotlivých tokenů a ignoruje veškeré zvláštnosti a specifika jazyka.

Translation memory (překladová paměť)

Jedná se o databázi termínů potřebnou pro činnost nástroje CAT tool. Obsahuje části textu přeložené v minulosti. CAT tool navrhuje (nebo automaticky zaměňuje) termíny z této databáze při procesu překládání a usnadňuje tím překladateli práci.

Trends (trendy)

Pojmem trend chápeme speciální vylepšení používané v diachronní analýze. Zkoumáme přitom frekvenci používání určitého slova (nebo slovního tvaru či jiného atributu) v závislosti na jejím vývoji v čase.

UMS

Vylepšení, které je dostupné uživatelům, kteří mají Sketch Engine nainstalován lokálně. Používá se administrativně pro správu uživatelů a korpusů.

User corpus (uživatelský korpus)

Tímto pojmem rozumíme korpus vlastnoručně vytvořený uživatelem. Uživatelé mohou vytvářet korpusy tak, že do Sketch Engine nahrají svá vlastní data, nebo je jeho prostřednictvím stáhnou z webu. Takto vzniklý korpus mohou pak dále sdílet s ostatními uživateli Sketch Engine.

Word form (slovní tvar)

Tímto termínem označujeme jednu konkrétní reprezentaci slova. Například u slovesa *jít* se jedná o každou z variací tohoto slova, a sice *šel, šli, jdeme, jdu* atd. Pokud v korpusu vyhledáváme podle slovních tvarů (například půjdeme), na výstupu se bude objevovat jen tento konkrétní slovní tvar (při rozlišování velkých a malých písmen vyhledávání slovních tvarů *Šel* a *šel* zobrazí různé výsledky)

Word list (seznam slov)

Souhrnný název pro mnohé výstupy vyhledávání v korpusech. Označujeme jím seznamy jak slov, tak i lemmat, slovních druhů, tagů, lempos a jiných. Současně je v seznamu slov zobrazena jejich frekvence zastoupení.

Word Sketch (Slovní profily)

Jednostránkový, automatický a korpusem odvozený souhrnný výsledek gramatického a kolokačního chování určitého slova nebo slovního tvaru.

Word Sketch grammar (Gramatika Word Sketch)

Zkratkou WSG rozumíme seznam pravidel definujících gramatické vztahy ve Word Sketch. WSG závisí na užitém jazyce, a proto nemůže být ten stejný set pravidel nikdy použit pro více jazyků. I různé korpusy ve stejném jazyce se mohou lišit ve WSG v nich použitých. Uživatelé Sketch Engine mohou vytvořit svá vlastní WSG, pokud jim žádná existující gramatika nevyhovuje. Pro korpusy v jazycích, které Sketch Engine nepodporuje, byla vytvořena univerzální gramatika, která se sestává pouze ze základních statistik týkajících se klíčových slov, a ta sice neobsahuje gramatiku konkrétního jazyka, ale dá se modifikovat.