CJDSL001 Korpusová lingvistika (3)

Klára Osolsobě

osolsobe@phil.muni.cz

Experimentální a počítačová lingvistika

O čem budeme mluvit v kurzu

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

Korpusová lingvistika – empirická disciplína - kde je možné užití korpusu

- Myšlenka korpusu ve strukturalistickém přístupu
- Korpus – dotazník – introspekce
- Volba/tvorba jazykového korpusu
- Typy korpusů
- Dostupné korpusy
- Problémové otázky
- Tvorba korpusu

Otázky, které je třeba vyřešit

- Je korpusový přístup vhodný pro náš výzkumný záměr ?
- Pokud je vhodný, existuje/existují vhodné korpus(y)?
- Pokud existují, existují i vhodné nástroje, které mi pomohou při plnění mého záměru?
- Pokud neexistují, je možné, aby vznikly?

Typologie korpusů

- Synchronní a diachronní (Je můj záměr orientován na současný jazyk, nebo na jazyk v diachronní perspektivě?)
- Psaný a mluvený (Je problém, který sleduji, typický pro psaný/mluvený jazyk, nebo je nezávislý na těchto aspektech?)
- Obecný a autorský (Je můj výzkumný problém orientován na jazyk jednoho/více autorů, nebo je otázka autorství textů druhořadá?)
- Specializovaný (Zaměřuji se na specifický problém, existuje k tomuto specifickému problému specificky zaměřený korpus?)
- Webový (Potřebuji k řešení svého výzkumného záměru především „velká data“?)
- Paralelní (Zajímá mě komparace více jazyků?)
- Srovnatelný (Zajímá mě komparace analogických jevů v různých jazycích?)

Otázky

- Máte přehled o dostupných korpusech jazyků, které alespoň částečně ovládáte?
- Korpusy v českém prostředí?
- Korpusy na MU?

Jak se orientovat v korpusech ČNK
https://www.korpus.cz/

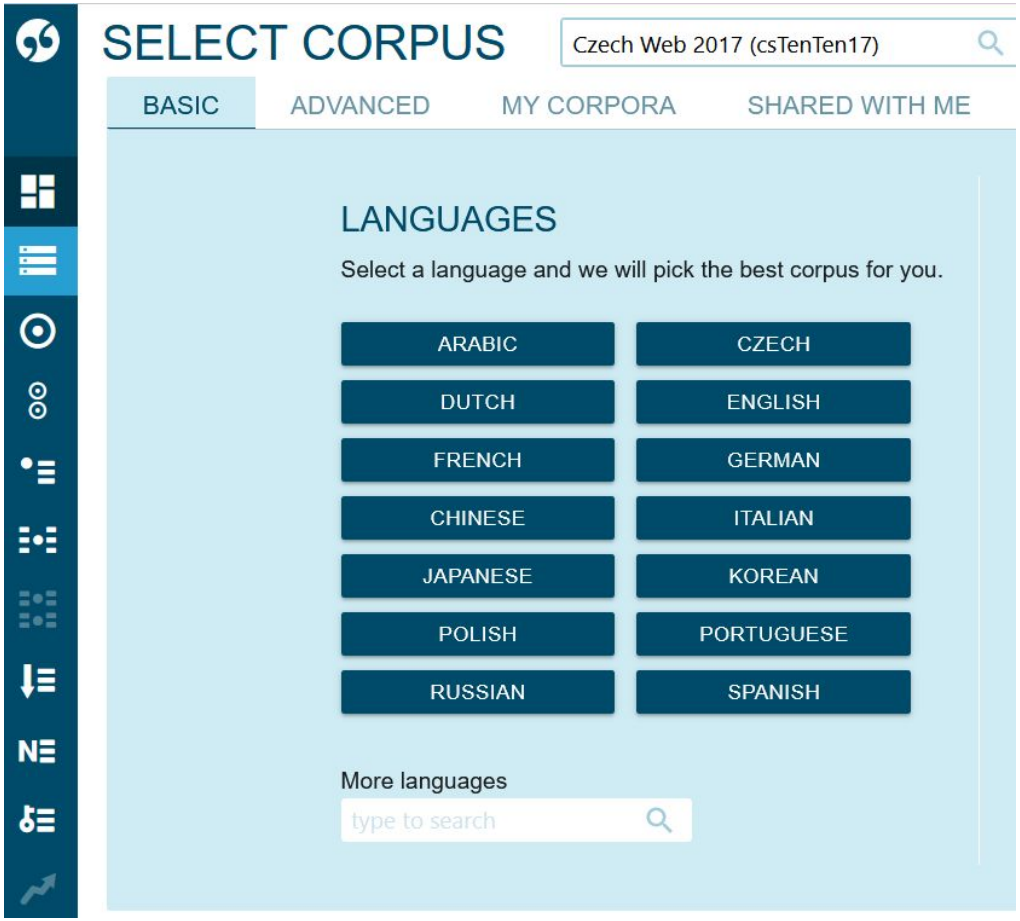
- akademický projekt 1994
- systematicky mapuje češtinu i další jazyky
- po bezplatné registraci otevřeny všem zájemcům (https://www.korpus.cz/signup)
- Korpusy – přehledně (https://wiki.korpus.cz/doku.php/cnk:uvod)

Jak číst jednotlivé charakteristiky a nad čím přemýšlet

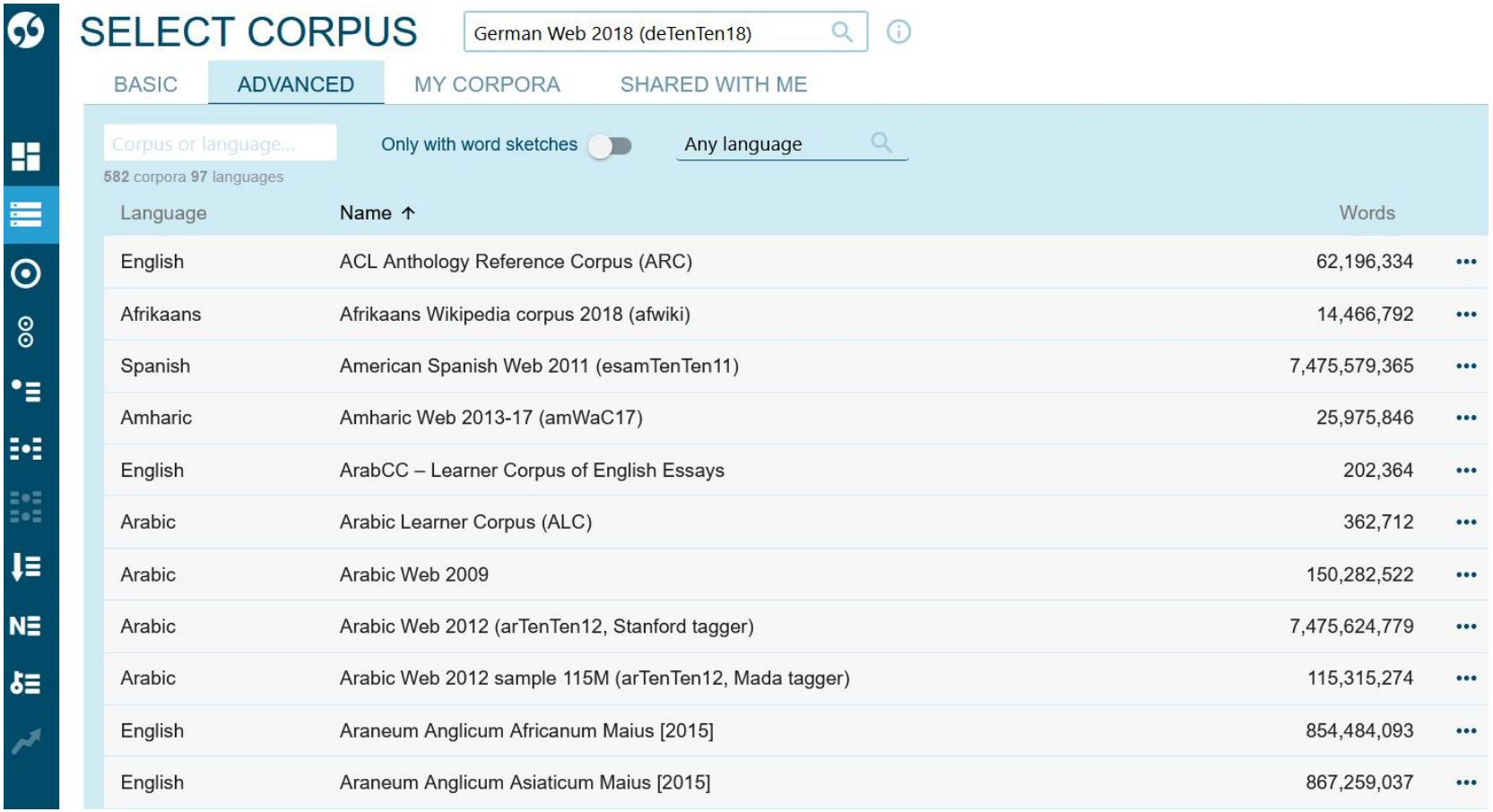
- Velikost korpusu (počet slov ve vztahu k tokenizace)
- Lemmatizace
- Morfologické značky
- Verzovaný korpus
- Referenční korpus
- Klasifikace textů – vnější anotace / metadata
- Citování korpusů

korpus	velikost (počet slov)	lemmatizace	morfologické značky	rok zveřejnění[1)]	charakteristika korpusu

Jak se orientovat v korpusech dostupných přes Sketch engine



Pokročilé


SELECT CORPUS
German Web 2018 (deTenTen18)
BASIC
ADVANCED
MY CORPORA
SHARED WITH ME
Corpus or language...
Only with word sketches
Any language
582 corpora 97 languages
Language
Name ↑
Words
English ACL Anthology Reference Corpus (ARC) 62,196,334
Afrikaans Afrikaans Wikipedia corpus 2018 (afwiki) 14,466,792
Spanish American Spanish Web 2011 (esamTenTen11) 7,475,579,365
Amharic Amharic Web 2013-17 (amWaC17) 25,975,846
English ArabCC – Learner Corpus of English Essays 202,364
Arabic Arabic Learner Corpus (ALC) 362,712
Arabic Arabic Web 2009 150,282,522
Arabic Arabic Web 2012 (arTenTen12, Stanford tagger) 7,475,624,779
Arabic Arabic Web 2012 sample 115M (arTenTen12, Mada tagger) 115,315,274
English Araneum Anglicum Africanum Maius [2015] 854,484,093
English Araneum Anglicum Asiaticum Maius [2015] 867,259,037

Podle klasifikace

Některé starší i novější projekty MU

- BMK (https://wiki.korpus.cz/doku.php/cnk:bmk)
- KSK (https://wiki.korpus.cz/doku.php/cnk:ksk-dopisy)
- Elektronická knihovna překladů anglických dramat (https://www.phil.muni.cz/kapradi/)
- UčKo

Korpusové nástroje

- Vyhledávání přes webové rozhraní KonText a Sketch Engine
- Funkce – vyhledávání, zobrazování, třídění, počítání frekvencí, ukládání, využití statistických měr
- Další korpusové nástroje: SyD, Morfio, Treq, KWords

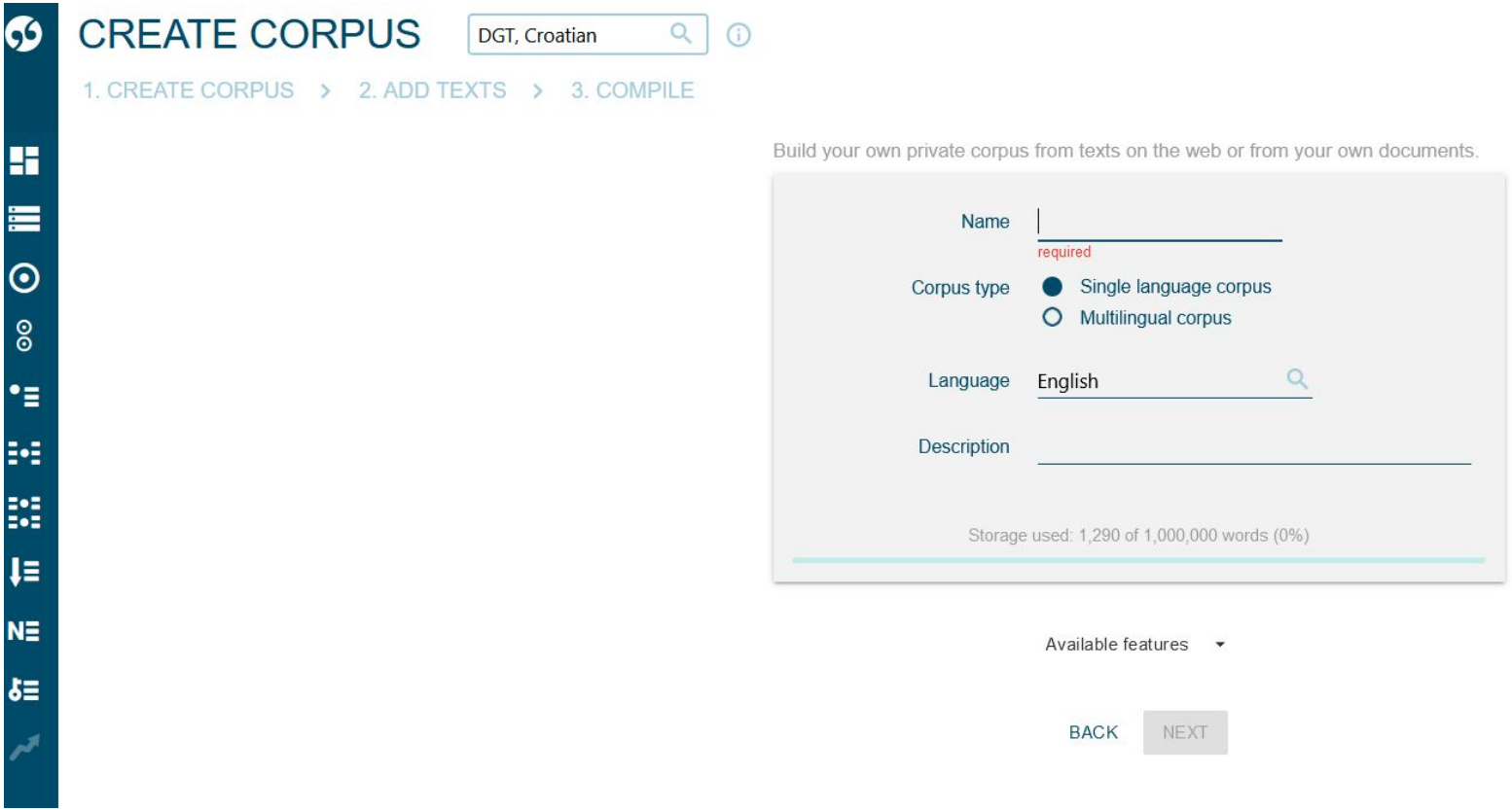
Vlastní korpus

- Tvorba subkorpusu z dostupného korpusu (https://www.korpus.cz/kontext/subcorpus/new?corpname=codit)
- Tvorba vlastního korpusu (https://app.sketchengine.eu/#ca-create?corpname=preloaded%2Fdgt__sh_hr)

Co je třeba řešit, chceme-li vytvořit vlastní korpus?

- Máme texty v elektronické podobě? Máme na ně právo?
- Jak dostat text do el. podoby?
- Čeká nás scanování a přepisy tetxů?
- Máme k dispozici vhodné OCR nástroje?
- Jsme dostatečně informováni o tom, jak udělat přepis?
- Jak zajistit kvalitu (konzistenci) přepisu? Mám prostředky na kontrolu přepisu?
- Jak budu zacházet s metadaty?

Tvorba korpusu a dostupné nástroje



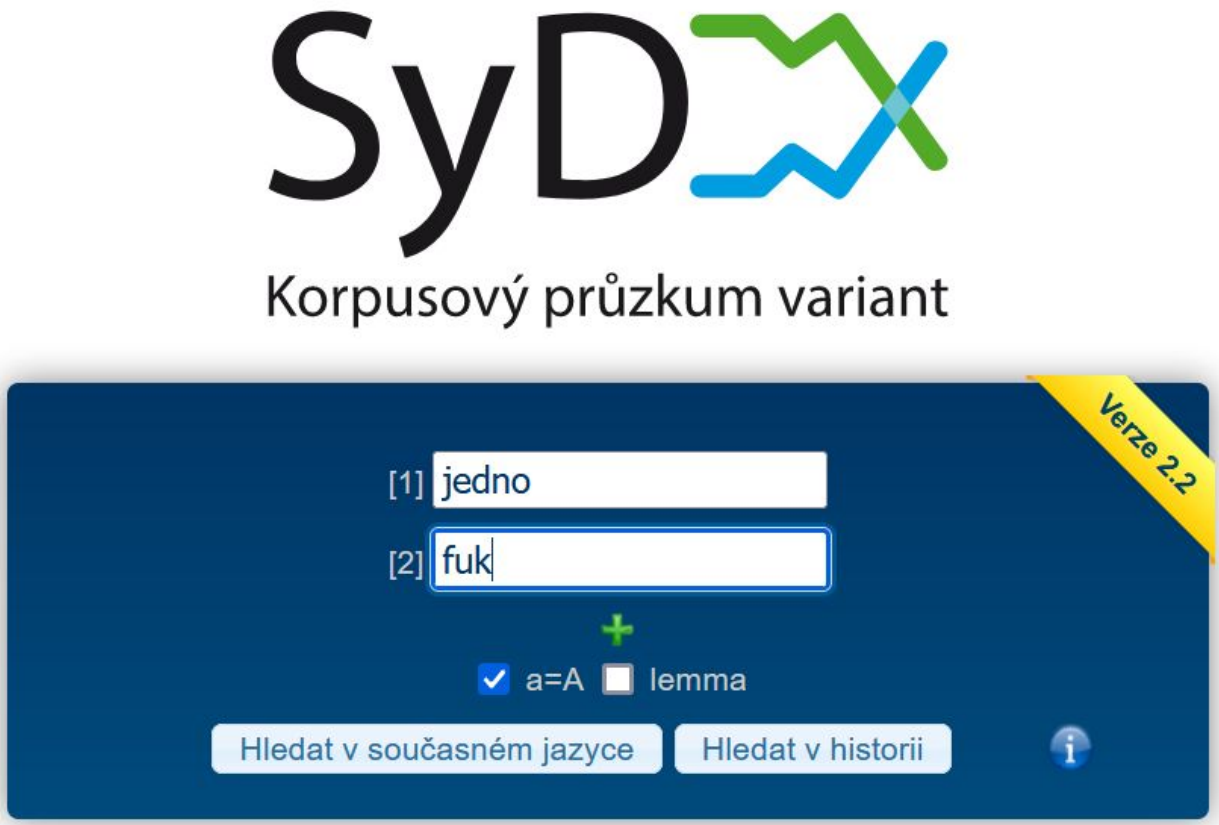
Závěr

- Máme k dispozici velké množství korpusů
- Korpusové nástroje jsou dostupné a mají množství funkcí
- Nástroje k vytvoření korpusu i pracoviště, na která je možno se obrátit existují
- Důležitost přípravných fází pro úspěšnost práce

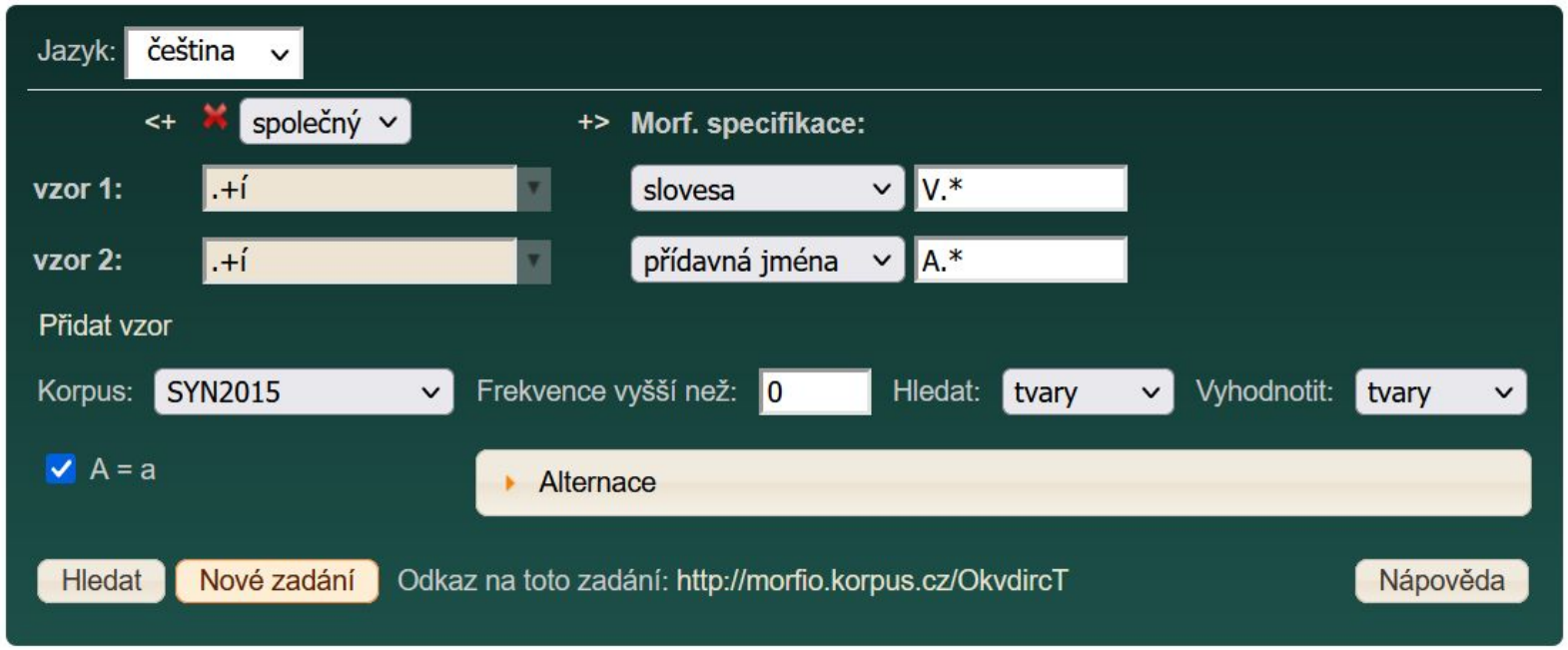
Vyzkoušejte

- V nástroji SyD porovnejte distribuci spojení *je třeba* a *je potřeba*.
- Jazykový humor a počítačové nástroje. Starý vtip: *„Koupil jsem si paštiku, bylo na ní napsáno ‚zaječí', a neječí a neječí."* Pomocí nástroje Morfio vyhledejte homonymní dvojice adjektivum/tvar slovesa podobné těm, které se objevují v uvedené anekdotě.
- Vyzkoušejte aplikaci KWords (vezměte svoji diplomovou práci/článek/referát, nahrajte ji/jej do nástroje KWord a podívejte se, nakolik se liší vámi sestavený seznam klíčových slov od seznamu vytvořeného automaticky uvedeným nástrojem. Udělejte totéž s originálem/překladem uměleckého textu. Vyzkoušejte obdobnou funkci v rozhraní Sketch engine
- Vyzkoušejte nástroj Treq (např. vyhledejte překlady vulgarismů, okazionalismů, podívejte se na falešné přátele)

Je mi to jedno/fuk.

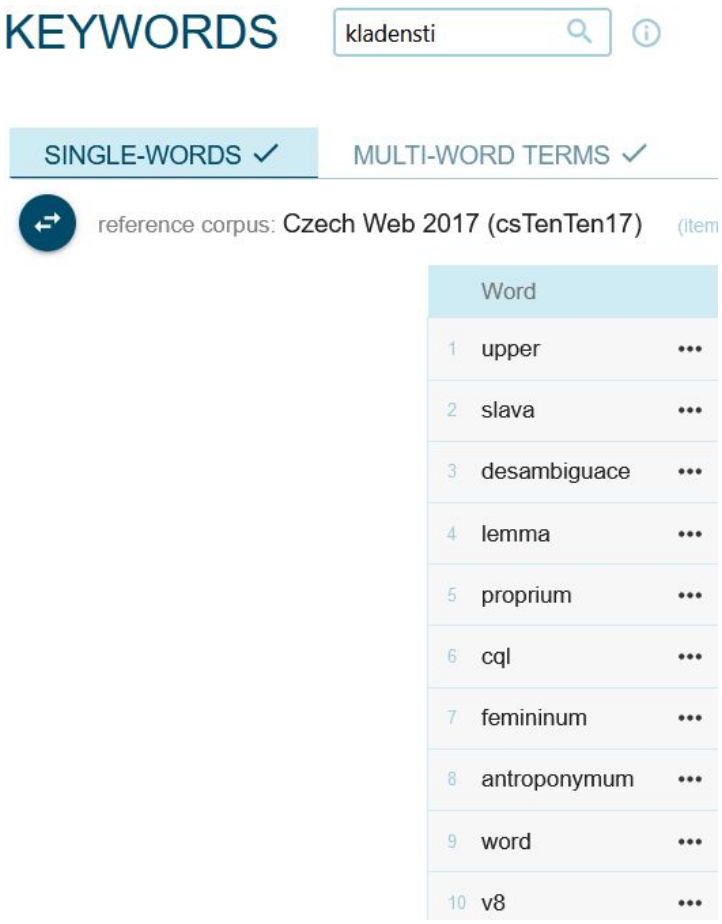


vlastnit/vlastní a další



antroponymum; oikonymum; zakončení na -slav/-slava; korpus; morfologické značkování; desambiguace

Funkce KEYWORDS ve Sketch engine



jak říci italsky *děvka?*

▲ Frekvence ▼	▲ Procenta ▼	▲ Čeština ▼	▲ Italština ▼
515	50.3	děvka	puttana
120	11.7	děvka	troia
80	7.8	děvka	stronzo
34	3.3	děvka	prostituta
33	3.2	děvka	sgualdrina
25	2.4	děvka	puttanella
19	1.9	děvka	Troia
16	1.6	děvka	cagna
14	1.4	děvka	zoccola
11	1.1	děvka	troietta
10	1.0	děvka	merda
8	0.8	děvka	donna
6	0.6	děvka	baldracca
5	0.5	děvka	bastardo
4	0.4	děvka	prostituto
4	0.4	děvka	stronzetta

Otázky

- Máte pro svou disertaci vybrán korpus, se kterým chcete pracovat?
- Budete tvořit vlastní korpus?
- Máte rozmyšlené, jak budou vypadat metadata?
- Zajímáte se o korpusově orientované konference a víte, že se chystá jedna na příští rok v Praze **(https://tt2022.ff.cuni.cz/calls-and-circulars/first-call-for-papers/)?**

Děkuji vám za pozornost