

# CJDSL001 Korpusová lingvistika (3)

Klára Osolsobě

[osolsobe@phil.muni.cz](mailto: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 <sup>1)</sup>	charakteristika korpusu
--------	-----------------------	-------------	---------------------	------------------------------	-------------------------



# Jak se orientovat v korpusech dostupných přes Sketch engine

**SELECT CORPUS** Czech Web 2017 (csTenTen17)

BASIC ADVANCED MY CORPORA SHARED WITH ME



### LANGUAGES

Select a language and we will pick the best corpus for you.


ARABIC	CZECH
DUTCH	ENGLISH
FRENCH	GERMAN
CHINESE	ITALIAN
JAPANESE	KOREAN
POLISH	PORTUGUESE
RUSSIAN	SPANISH

More languages  
type to search

# Pokročilé

**SELECT CORPUS**   

BASIC **ADVANCED** MY CORPORA SHARED WITH ME

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

## CORPUS CATEGORY

ALL ⓘ 582

RECENT ⓘ 0

MY CORPORA ⓘ 5

SHARED WITH ME ⓘ 9

FEATURED ⓘ 15

GENERAL PURPOSE ⓘ 316

WEB ⓘ 262

NON-WEB ⓘ 320

PARALLEL ⓘ 148

SPOKEN ⓘ 53

SPECIALIZED ⓘ 223

DIACHRONIC ⓘ 133

MULTIMEDIA ⓘ 1

LEARNER ⓘ 5

ERROR-ANNOTATED ⓘ 3

GRAMMAR DEVELOPMENT ⓘ 0

# 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 měř
- 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%2Fdg\\_\\_sh\\_hr](https://app.sketchengine.eu/#ca-create?corpname=preloaded%2Fdg__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

## CREATE CORPUS

DGT, Croatian

1. CREATE CORPUS > 2. ADD TEXTS > 3. COMPILE

Build your own private corpus from texts on the web or from your own documents.

Name   
required

Corpus type  Single language corpus  
 Multilingual corpus

Language

Description

Storage used: 1,290 of 1,000,000 words (0%)

Available features



# 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

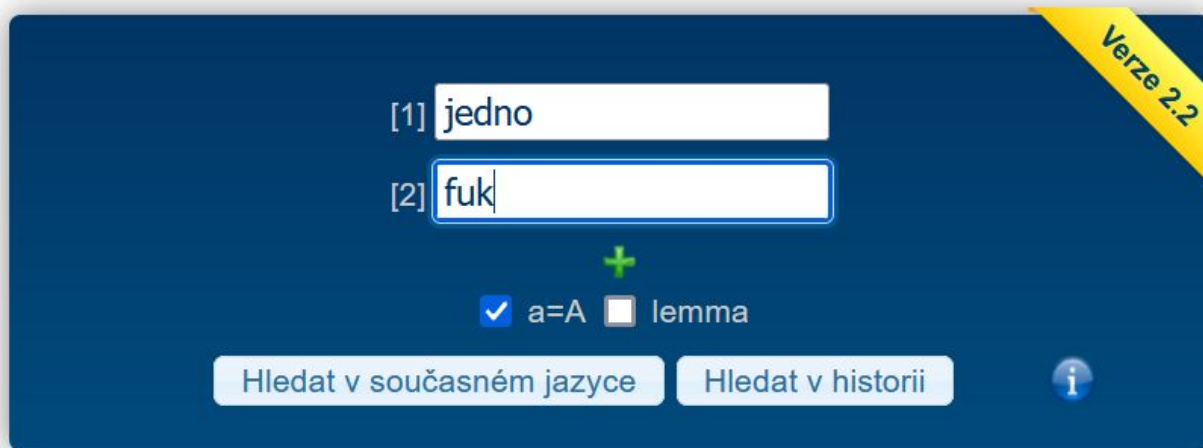
# Vyzkoušejte

- 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.*



Korpusový průzkum variant



The screenshot shows the SyD search interface on a dark blue background. At the top right, a yellow ribbon banner reads 'Verze 2.2'. Below it, there are two search input fields: the first is labeled '[1]' and contains the text 'jedno'; the second is labeled '[2]' and contains 'fuk'. A green plus sign is centered between the two fields. Below the plus sign, there are two checkboxes: the first is checked and labeled 'a=A', and the second is unchecked and labeled 'lemma'. At the bottom, there are two light blue buttons: 'Hledat v současném jazyce' and 'Hledat v historii'. To the right of these buttons is a small blue circular icon with a white lowercase 'i' inside, representing an information or help button.

# *vlastnit/vlastní a další*



Jazyk: čeština ▾

<+ ✖ společný ▾ +> Morf. specifikace:

vzor 1: .+í ▾ slovesa ▾ V.\*

vzor 2: .+í ▾ přídavná jména ▾ A.\*

Přidat vzor

Korpus: SYN2015 ▾ Frekvence vyšší než: 0 Hledat: tvary ▾ Vyhodnotit: tvary ▾

A = a



▶ Alternace

Hledat Nové zadání Odkaz na toto zadání: <http://morfio.korpus.cz/Okvdirct> Nápověda


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



# Funkce KEYWORDS ve Sketch engine

KEYWORDS   

SINGLE-WORDS ✓ MULTI-WORD TERMS ✓

 reference corpus: Czech Web 2017 (csTenTen17) (item

	Word	
1	upper	...
2	slava	...
3	desambiguace	...
4	lemma	...
5	proprium	...
6	cql	...
7	femininum	...
8	antroponymum	...
9	word	...
10	v8	...

# jak říci italsky *děvka*?

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

# 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