CORE147

DIGITÁLNÍ DATA V HUMANITNÍCH A SOCIÁLNÍCH

Zuzana Nevěřilová
Hana Žižková

2024/25

TEXT

What does it mean, studying world literature?
How do we do it?
I work on West European narrative between 1790 and 1930,
and already feel like a charlatan outside of Britain or France.

World literature?

Reading "**more**" seems hardly to be the solution.

TEXT

"I work on West European narrative, etc. …"
Not really,
I work on its **canonical fraction**,
which is not even
one per cent of published literature.

"The great unread"
Margaret Cohen

Franco Moretti: Conjectures on World Literature, 2000

TEXTOVÁ DATA

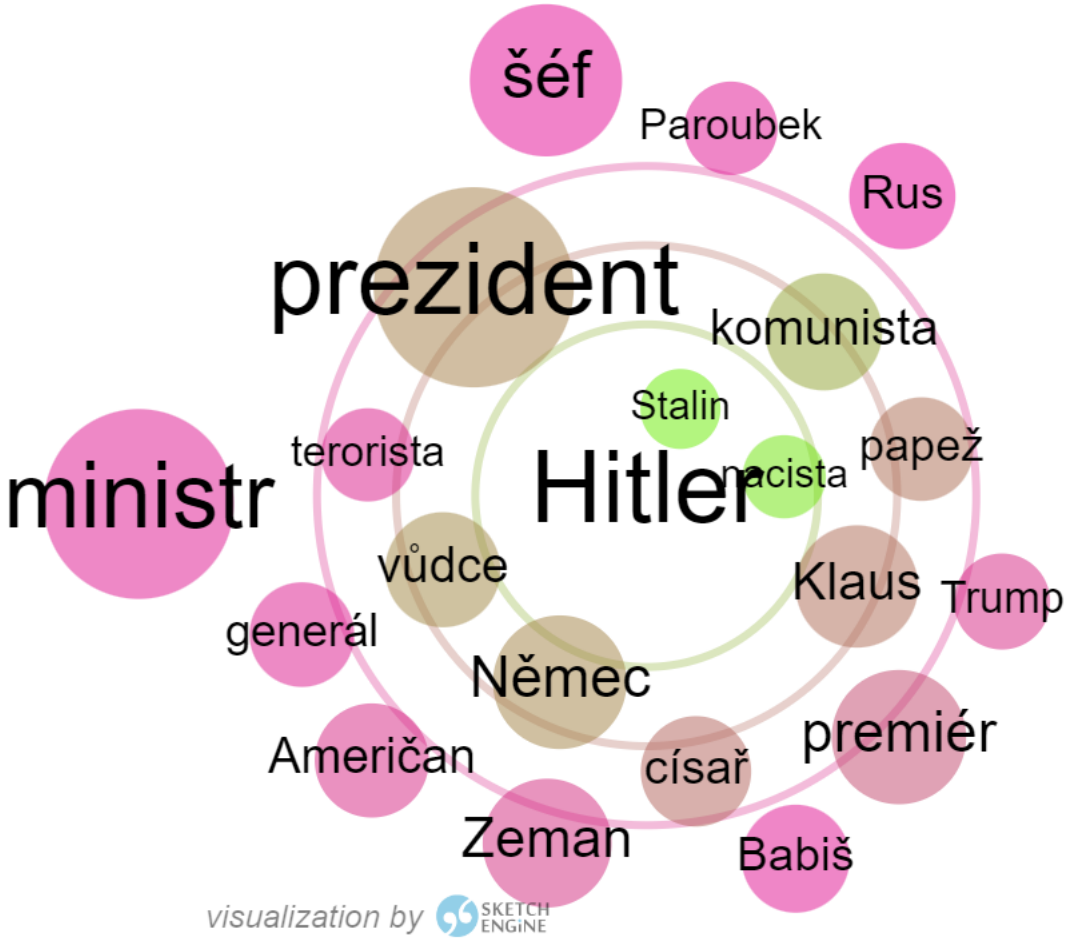
- Kdo to bude číst?
- Všechno? Nikdo. Něco? Někdo.

Close reading	Distant reading
kvalitativní výzkum	kvantitativní výzkum
zpracujeme menší vzorek	zpracujeme větší vzorek
vidíme více detailů	vidíme méně detailů
statistické metody nelze použít nebo jen omezeně	statistické metody mohou být velmi užitečné

TEXTOVÁ DATA

- **Korpus** = soubor textů, anotací a metadat
- **Texty** = souvislé, vyprodukované lidmi, případně jako výsledek přepisu mluveného slova
- **Anotace** = označkování jako výsledek analýzy textu, např. segmentace na slova a věty
- **Metadata** = informace o jednotlivých dokumentech

zmrazit ús
esvědčeni
rů! Třicátél
měnit. O tu
marskou r
e. A při výs
a Háchy u
u s Němci
čně. Křest



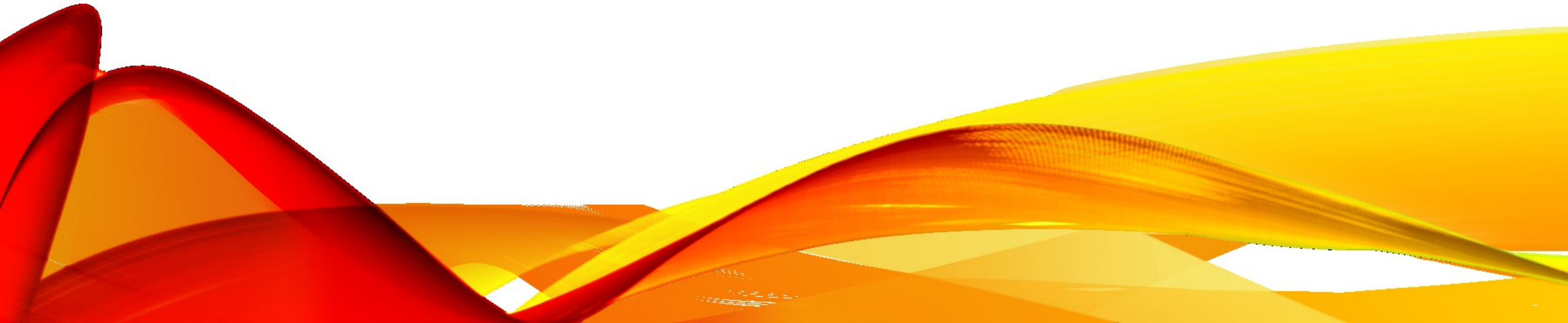
ZPRACOVÁNÍ PŘIROZENÉHO JAZYKA
NATURAL LANGUAGE PROCESSING

Přirozené jazyky a formální jazyky

- fonetická a fonologická
- morfologická
- gramaticko-lexikální
- syntaktická
- sémantická
- pragmatická (mimo jazyk)
- logická (mimo jazyk)
- znalosti o světě (mimo jazyk)

KORPUSOVÝ MANAŽER

Corpus manager

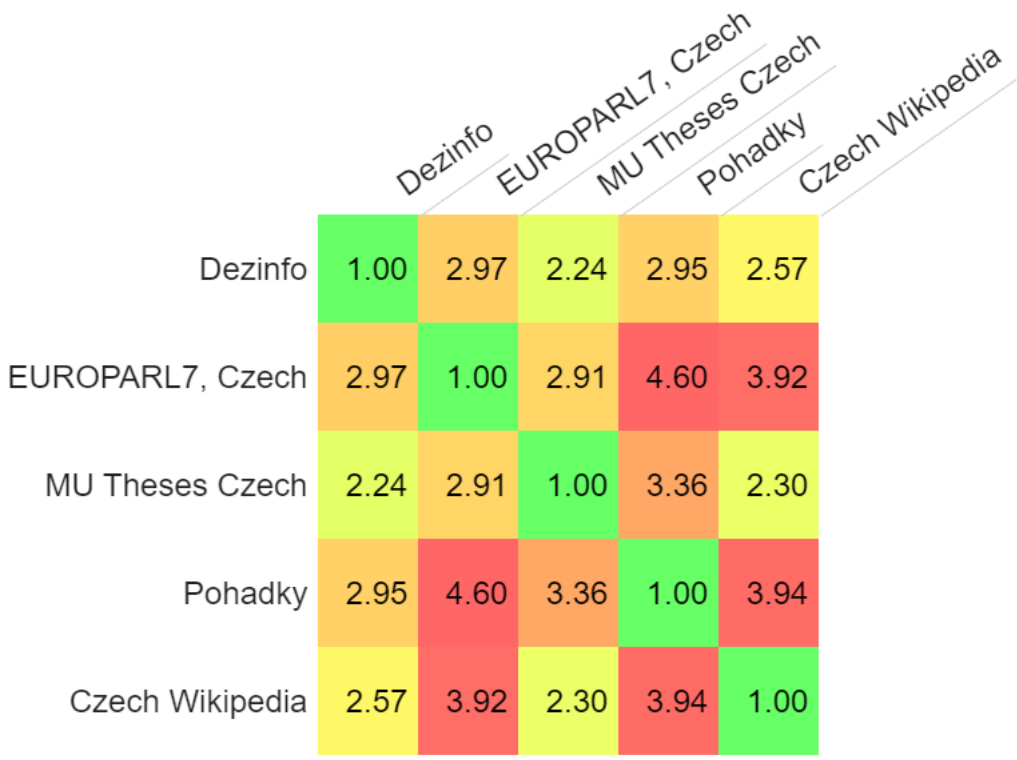


PROČ SPECIÁLNÍ SOFTWARE?

- Hledání ve velkém množství textu (největší korpusy mají miliardy slov)
- Rychlost vyhledávání
- Komplexní dotazy
  - Dotazovací jazyk CQL
  - Regulární výrazy
- Texty prošly anotací
- Texty obsahují metadata, podle kterých lze také hledat
- Výsledky hledání lze ukládat a dál s nimi pracovat
  - Kolekce
  - Vizualizace

KORPUSOVÝ MANAŽER
VÝBĚR KORPUSU

- Jednojazyčné korpusy
- Paralelní
  - zarovnané
- Korpusy autora
- Mluvené korpusy
- Korpusy podle žánru
- Korpusy podle subkultury
  - Tweety, diskuzní fóra
- Webový korpus
- Vyvážený korpus



KORPUSOVÝ MANAŽER DOTAZ

- Konkordance (slovo „ve svém přirozeném prostředí“)
  - Jednoduchý dotaz
  - Dotazovací jazyk CQL
- Seznam slov (podle frekvence)
  - Stop slova
- Kolokace (slova, která jsou často spolu)
- N-gramy
- Slovní profil (word sketch) – sumarizace konkordancí
- Slovo v kostce

DOTAZOVACÍ JAZYK CQL

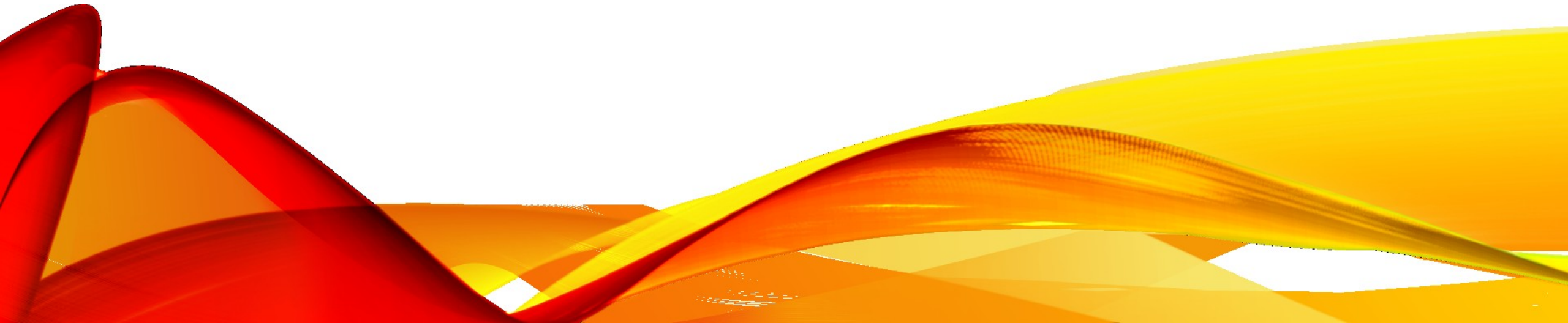
- Pozice (token)
- Atribut
- Hodnota

slovo	a	maji	tam	cedule	ňáký	,	kde	prodávají	jabka
lemma	a	mít	tam	cedule	nějaký	,	kde	prodávat	jablko
POS	J	V	ADV	N	A		J	V	N

```
[word="[Jj]abl?[k].*"]

[lemma="prodávat"] [tag="N.*"]
```

UKÁZKY KORPUSOVÝCH MANAŽERŮ



VYHLEDÁVÁNÍ V KORPUSU

- Český národní korpus https://www.korpus.cz/kontext/
- SketchEngine https://ske.fi.muni.cz/
- ParlaMint https://lindat.mff.cuni.cz/services/teitok/parlamint-41/

DALŠÍ ČTENÍ

- Franco Moretti: The Slaughterhouse of Literature. *Modern Language Quarterly*. 1 March 2000; 61 (1): 207–228. doi: https://doi.org/10.1215/00267929-61-1-207. https://msu.edu/course/eng/487/snapshot.afs/johnsen/61.1moretti.pdf
- Franco Moretti: Conjectures on World Literature. New Left Review. 1. 54-68. 2014. https://newleftreview.org/issues/ii1/articles/franco-moretti-conjectures-on-world-literature
- Adam Kilgarriff: Structured bibliography. Sketch Engine. 2015. https://www.sketchengine.eu/adam-kilgarriff-structured-bibliography/
- CQL – Corpus Query Language. Sketch Engine. Lexical Computing CZ s.r.o. https://www.sketchengine.eu/documentation/corpus-querying/