Lexical Computing

Vojtěch Kovář

Brighton, UK & Brno, CZ

Setkání SPP FI MU
listopad 2017

Lexical Computing

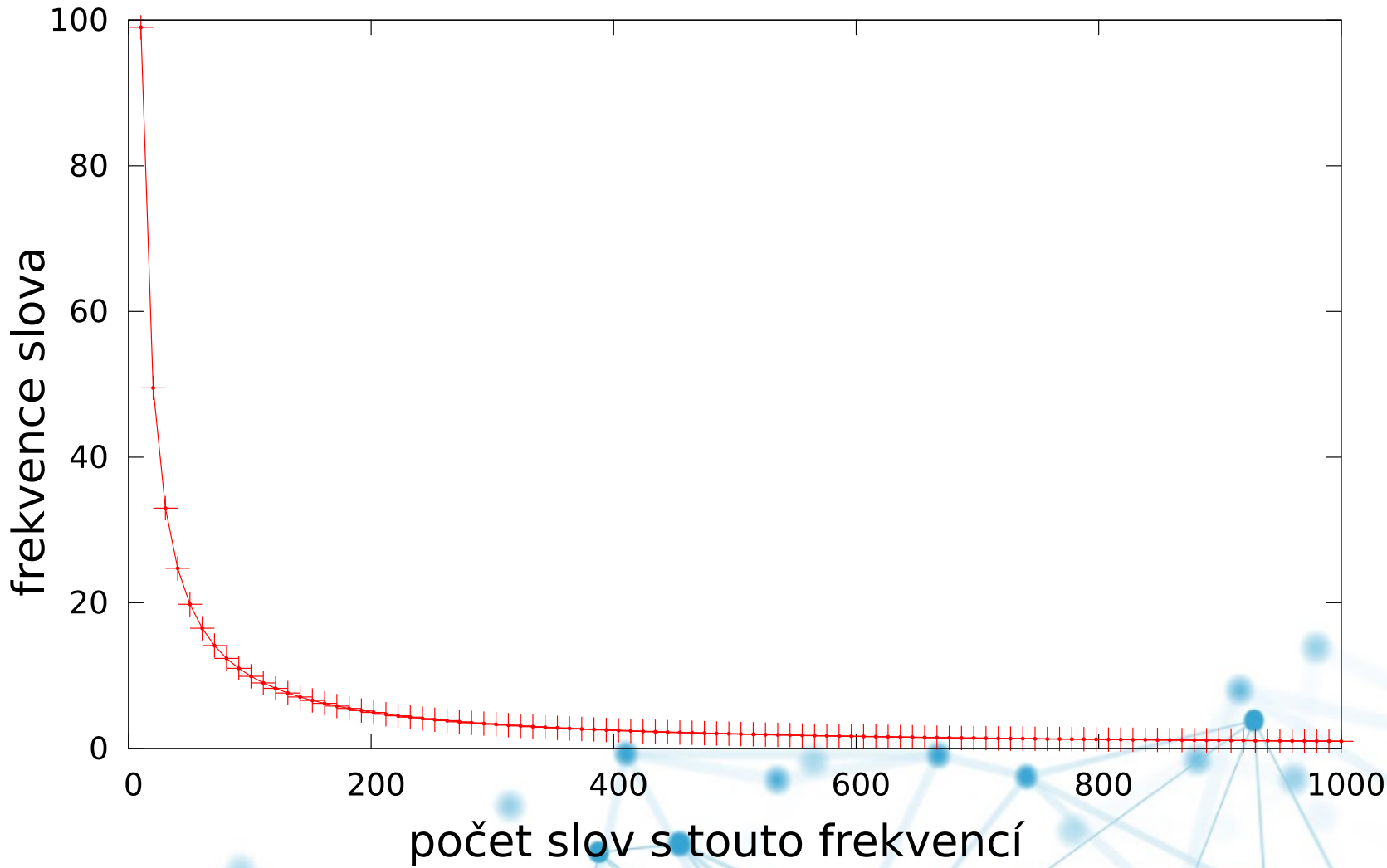
- výzkumná firma založená v Británii v roce 2003
- jazykové technologie, zejména korpusová lingvistika a počítačová lexikografie
- významný podíl činnosti tvoří vlastní výzkum
- partner na FI: Centrum zpracování přirozeného jazyka

Co je to korpus

Textový korpus

- rozsáhlý soubor textů
- často jazykově označkovaný
  základní tvary, slovní druhy, morfologické značky
- všeobecné i vysoce specializované (oborově, jazykově)
- obsahuje aktuální užití jazyka
- pro reprezentativnost musí být opravdu velký (10^9+ slov)

Zipfovo rozložení



Využití korpusů

- komerční:
  - počítačová lexikografie – tvorba slovníků
  - informační systémy, analytické nástroje
  - inteligentní (mobilní) aplikace
- akademické:
  - univerzity pro výzkum a výuku jazyků
  - studenti jazyků

Uživatelé


CAMBRIDGE
UNIVERSITY PRESS


Cornelsen


KDICTIONARIES


OXFORD
UNIVERSITY PRESS


Collins


Dictionary.com


LE ROBERT


小学館
SHOGAKUKAN

Uživatelé

- lexikografové
- vědci
- učitelé
- studenti
- překladatelé
- terminologové
- textaři, kreativci

Sketch Engine

- hlavní produkt LCL
- webová aplikace
- nástroj na zpracování a zkoumání textových korpusů
- jazykově nezávislý, velké množství funkcí pro různá použití
- stovky korpusů, desítky tisíc uživatelů
- pro kohokoli z MU volný přístup na http://ske.fi.muni.cz

Sketch Engine

více než **600 korpusů** pro **89 jazyků**:

- 100+ korpusů větších než 100M pozic
- 30+ korpusů větších než 1G pozic
  - od roku 2010 korpusy TenTen (10^{10})
- 60+ jazyků s anotací slovních tvarů
- 50+ jazyků s anotací slovních profilů (word sketches)
- vícejazyčné korpusy

Word sketch

water *(noun)* British National Corpus freq = 34246 (305.3 per million)

modifier	9591	1.1
hot	665	10.17
drinking	352	9.97
cold	459	9.63
boiling	237	9.58
fresh	231	8.81
mineral	173	8.76
running	145	8.74
Thames	140	8.54

object_of	5126	1.6
pump	92	8.82
pour	139	8.74
drink	133	8.55
heat	72	8.43
boil	55	8.12
tread	43	7.88
fetch	41	7.51
splash	29	7.26

subject_of	2835	1.7
flow	113	9.29
drip	36	8.33
seep	30	8.2
gush	23	7.94
boil	25	7.62
cascade	17	7.48
swirl	18	7.45
lap	17	7.43

Formy spolupráce LCL a CZPJ

- bakalářské, diplomové, disertační práce pod vedením/na náměty LCL
- více než 40 obhájených BP a DP, 2 sponzorované PhD
- práce na společných grantech (EU projekt PRESEMT)
- vytváření, úpravy a anotace korpusů
- vývoj algoritmů nad korpusy

Řešené problémy z pohledu informatika

- jak rychle zpracovat desítky TB dat
- jak taková data efektivně indexovat
- jak v indexech o desítkách GB hledat ve zlomcích vteřiny
- jak spočítat podobnost matic 1M×1M
- ...

Vytvořené webové korpusy

enClueWeb09:

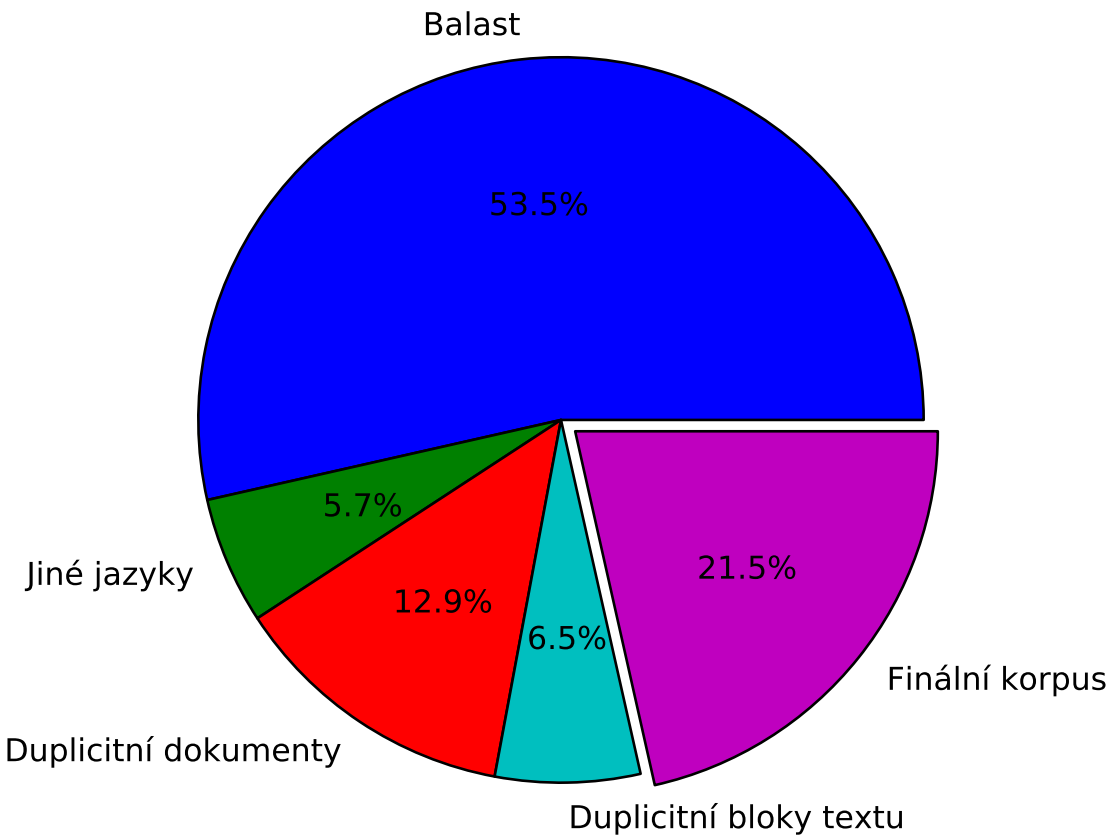
	Size [mil]
total tokens	81 990
alphanumeric	70 330
numbers	1 485
punctuation	9 849
documents	138

korpusy řady TenTen:

jazyk	cz	de	en	es	fr	it
mld. tokenů	5,4	2,8	33,1	9,8	12,4	3,1
jazyk	jp	pt	ru	sk	zh	
mld. tokenů	11,1	0,9	20,2	0,9	2,1	

Zpracování EnTenTen

původní velikost 14,6 mld slov, výsledek 3,2 mld slov



SpiderLing

- webový crawler pro vytváření velkých korpusů
- zaměřuje se na webové domény bohaté na text
- paralelní zpracování získaných dat
- napsán v Pythonu

Závěr

Lexical Computing

- korpusová lingvistika a jazykové technologie

Kontakt

- milos.jakubicek@sketchengine.co.uk
- vojtech.kovar@sketchengine.co.uk
- místnost S214