Setkání SPP FI

Miloš Jakubíček

LEXICAL COMPUTING

Brighton, UK & Brno, CZ

Setkání SPP FI MU
6. prosince 2018

Lexical Computing

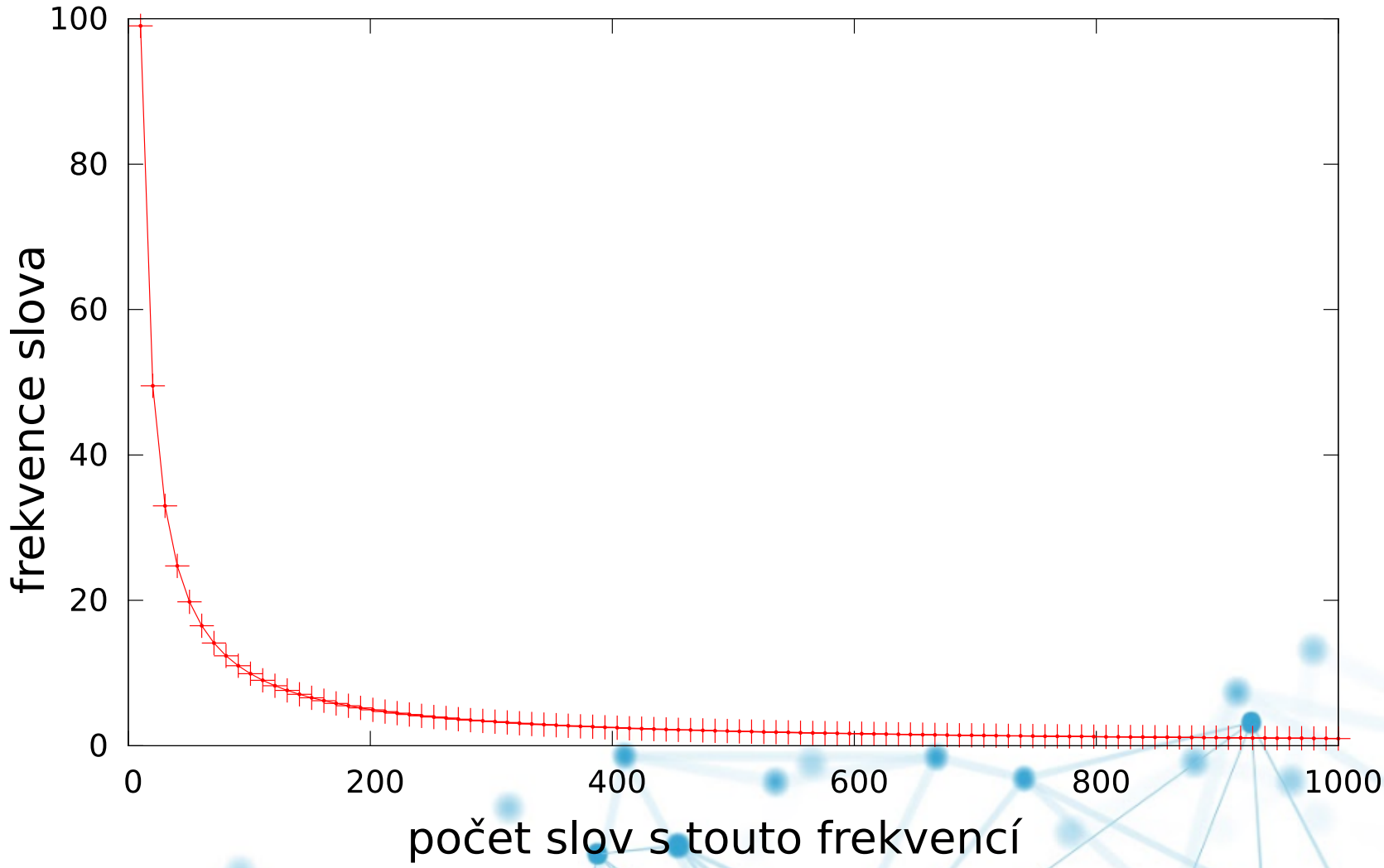
- výzkumná firma založená v Británii v roce 2003
- působí v oblastech korpusové lingvistiky a počítačové lexikografie
- významný podíl činnosti tvoří vlastní výzkum
- partner na FI: Centrum zpracování přirozeného jazyka

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

k prosinci 2018 více než **500 korpusů** pro **100+ 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
- 42 jazyků s anotací slovních profilů
- vícejazyčné korpusy

Formy spolupráce LCL a CZPJ

- bakalářské, diplomové, disertační práce pod vedením/na náměty LCL
- k prosinci 2018 40 obhájených BP, 11 DP, 4 PhD, nyní 3 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 efektivně distribuovat 300TB pro vyhledávání
- jak spočítat podobnost matic 1M$\times$1M
- ...

Vytvořené webové korpusy

enClueWeb09:

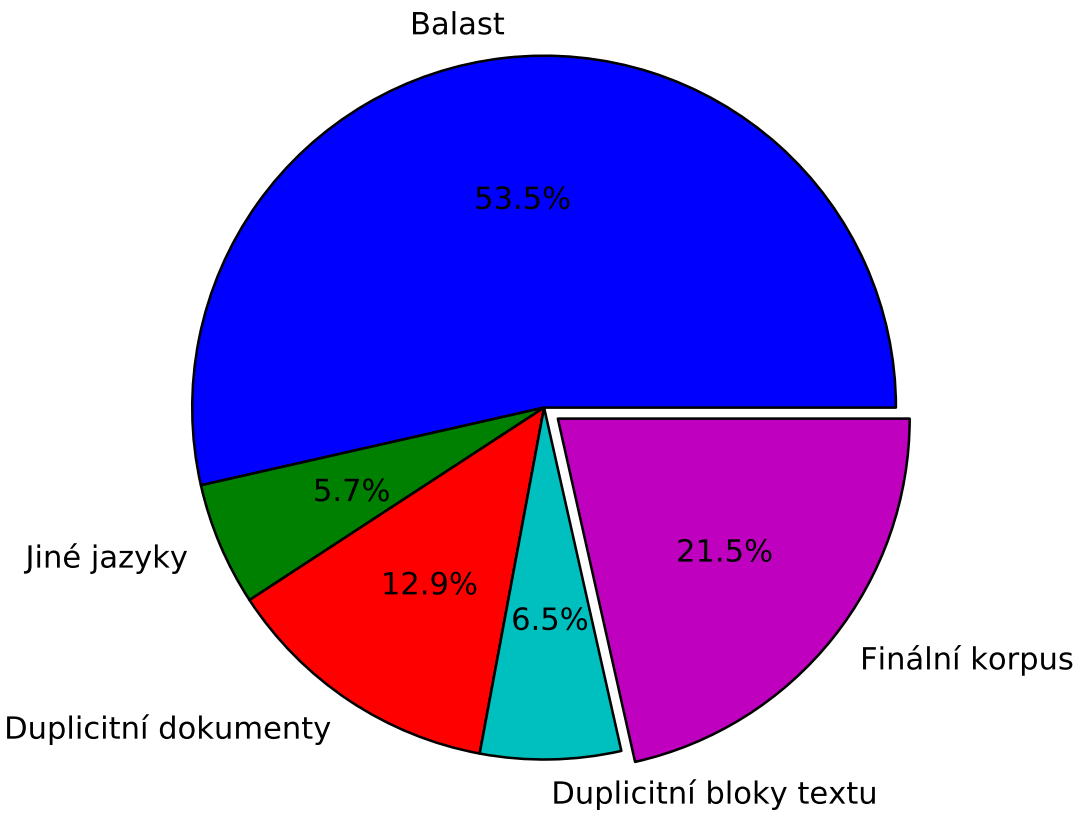
	Size [mil]
total tokens	81 990
alphanumeric	70 330
numbers	1 485
punctuation	9 849
documents	138

korpusy řady TenTen:

jazyk	cz	de	en	es	fr	it
mld. tokenů	5,4	2,8	13,0	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 a počítačová lingvistika

Kontakt

- milos.jakubicek@sketchengine.co.uk
- místnost S211