

# Lexical Computing Ltd.

Miloš Jakubíček

Setkání SPP FI MU  
6. května 2015

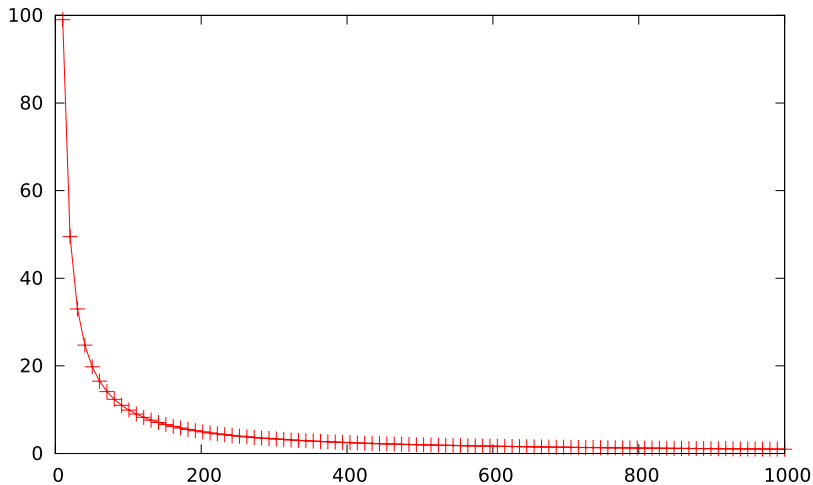
- menší britská firma
- založil Adam Kilgarriff v roce 2003
- působí v oblastech korpusové a počítačové lingvistiky
- významný podíl činnosti tvoří vlastní výzkum
- partner na FI: Centrum zpracování přirozeného jazyka



## Textový korpus

- rozsáhlý soubor textů
- často jazykově označovaný  
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



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

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

k prosinci 2014 více než **400 korpusů** pro **70 jazyků**:

- 100+ korpusů větších než 100M pozic
- 30+ korpusů větších než 1G pozic
  - od roku 2010 korpusy TenTen ( $10^{10}$ )
- 56 jazyků s anotací slovních tvarů
- 36 jazyků s anotací slovních profilů

- bakalářské, diplomové, disertační práce pod vedením/na náměty LCL
- práce na společných grantech (EU projekt PRESEMT)
- vytváření korpusů
- úpravy a anotace korpusů
- vývoj algoritmů nad korpusy



- bakalářské, diplomové, disertační práce pod vedením/na náměty LCL
  - práce na společných grantech (EU projekt PRESEMT)
  - vytváření korpusů
  - úpravy a anotace korpusů
  - vývoj algoritmů nad korpusy
- 
- není potřeba 4 měsíce školení ani certifikáty na složité systémy

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

# Vytvořené webové korpusy

## enClueWeb09:

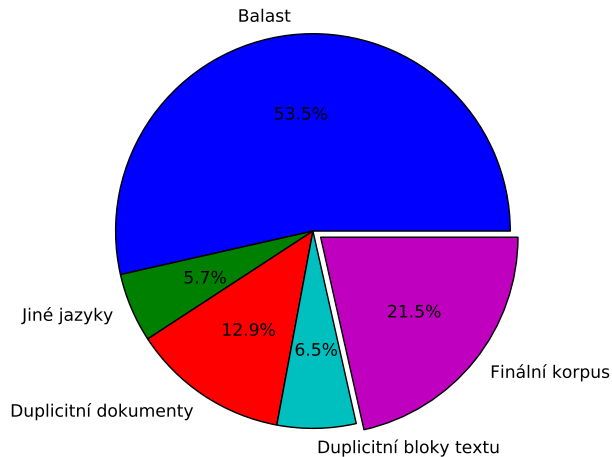
	Size [mil]
total tokens	81 990
alphanumeric	70 330
numbers	1 485
punctuation	9 849
documents	138

## 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



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

## Lexical Computing

- korpusová a počítačová lingvistika

## Kontakt

- [milos.jakubicek@sketchengine.co.uk](mailto:milos.jakubicek@sketchengine.co.uk)
- stánek v přízemí