Lexical Computing Ltd.

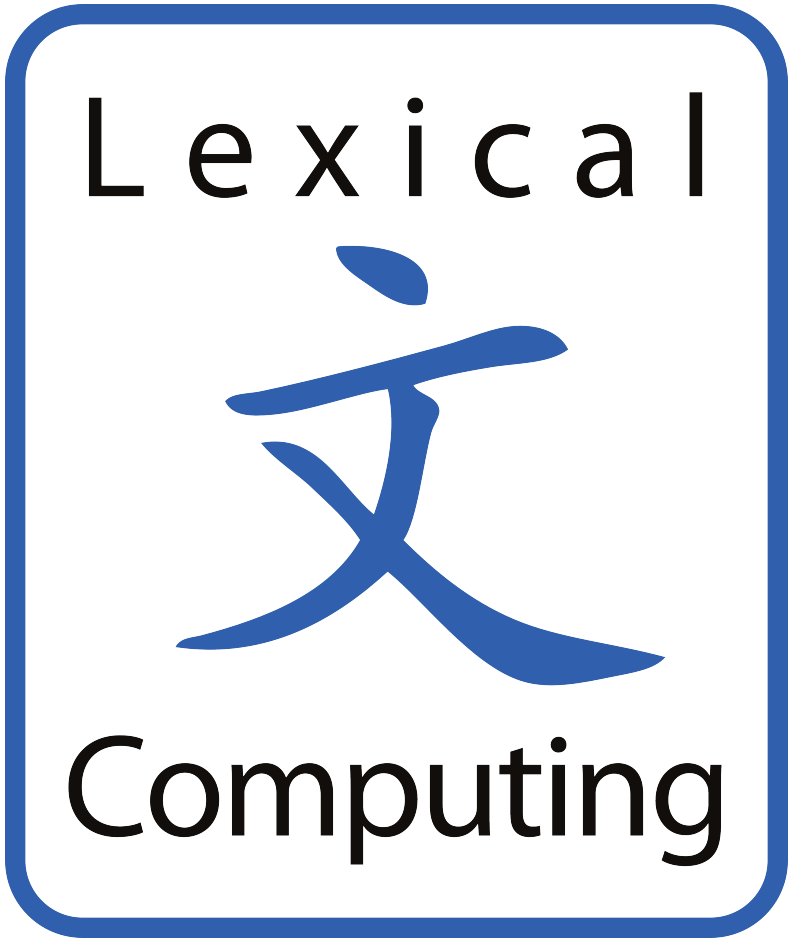
NLP Centre FI MU

Pavel Rychlý

December 6, 2012

Lexical Computing Ltd.

- menší britská firma
- založil Adam Kilgarriff v roce 2003
- působí v oblastech korpusové a počítačové lingvistiky
- významný podíl činnosti: výzkum
- moto: *corpora for all*



Sketch Engine

- 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

Co je to korpus

Textový korpus

- rozsáhlý soubor textů
- často jazykově označkovaný
  základní tvary, slovní druhy, morfologické značky
- všeobecné i vysoce specializované
  např. korespondence Bedřicha Smetany

Co je to korpus

Textový korpus

- rozsáhlý soubor textů
- často jazykově označkovaný
  základní tvary, slovní druhy, morfologické značky
- všeobecné i vysoce specializované
  např. korespondence Bedřicha Smetany
- obsahuje užití jazyka

Využití korpusů

- lingvisté, lexikografové
- studenti jazyků
- jazykový výzkum
- koncové aplikace (T9 a podobné)

Řešené problémy

- jak rychle zpracovat TB dat
- jak GB indexech hledat ve zlomcích vteřiny
- jak spočítat podobností matici 1M$\times$1M
- ...

Formy spolupráce LCL a Centra NLP

- 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

Formy spolupráce LCL a Centra NLP

- 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 začít pracovat pro LCL: práce v Centru NLP

Vytvořené webové korpusy

enClueWeb09:

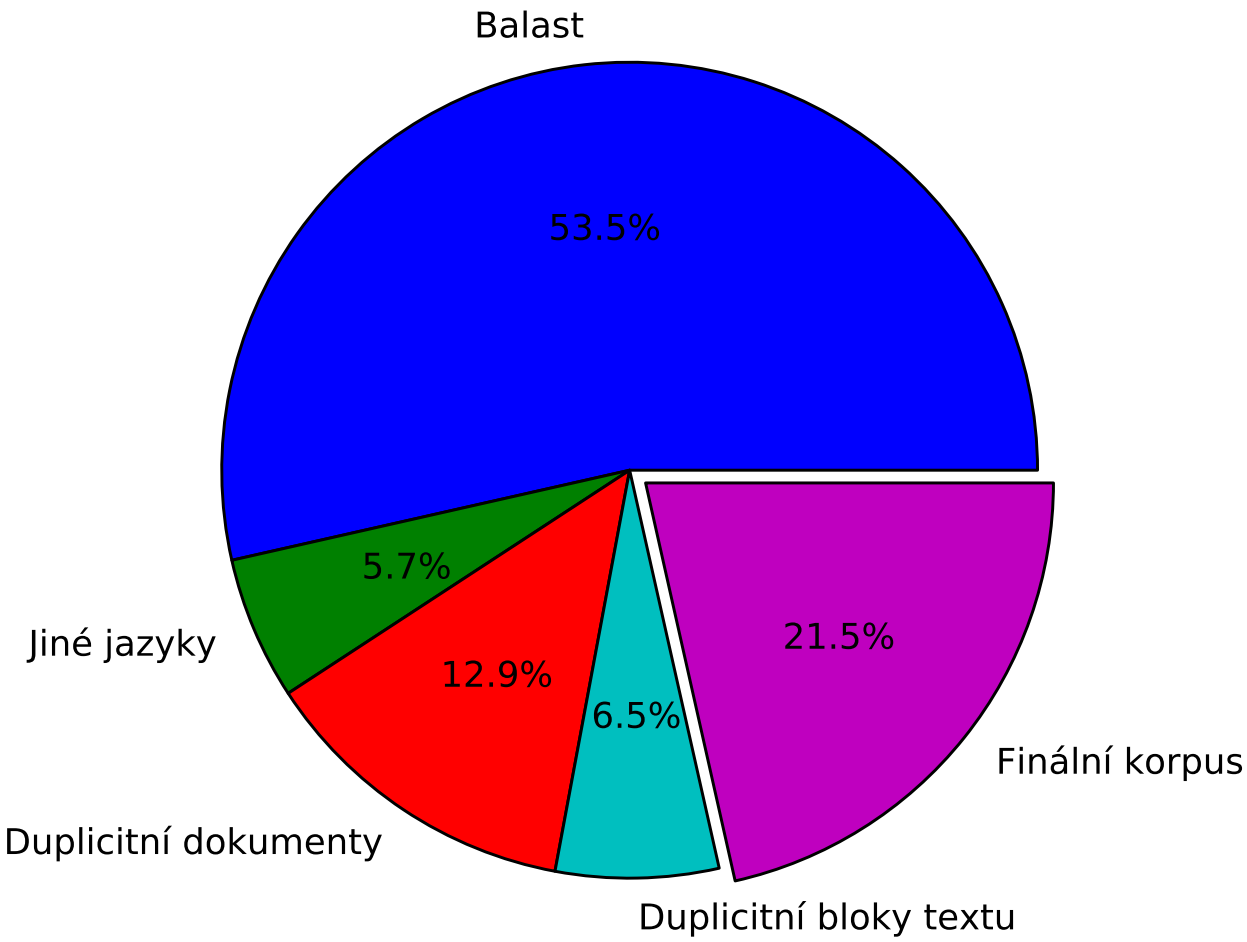
	Size [mil]
total tokens	81 990
alphanumeric	70 330
numbers	1 485
punctuation	9 849
documents	138

korpusy řady TenTen:

jazyk	cz	de	en	es	fr	it	jp	pt	ru
mld. tokenů	5,4	2,8	13,0	9,8	12,4	3,1	11,1	0,9	20,2

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
- od září 2011 v testovacím provozu