

Lexical Computing Ltd.

NLP Centre FI MU

Pavel Rychlý

April 26, 2012

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

- 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

Textový korpus

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

Textový korpus

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

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

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

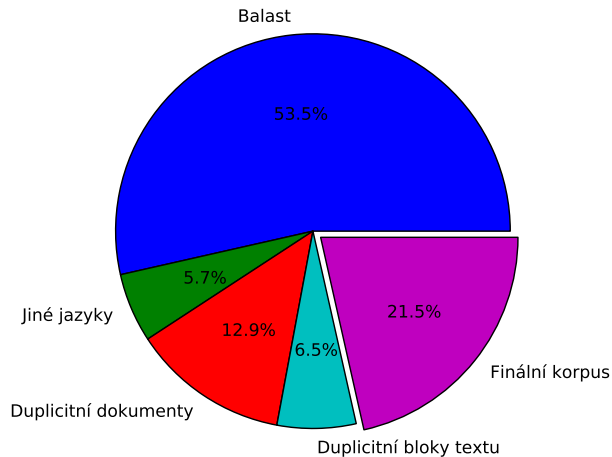
Korpusy řady TenTen

- cílová velikost korpusu $10^{10}=10$ mld. slov
- různé jazyky

jazyk	cz	de	en	es	it	pt	sk	zh
mil.	5828	1151	3268	2538	3076	948	876	2106

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
- od září 2011 v testovacím provozu