Korpusy textů a jejich využití

Pavel Rychlý, Aleš Horák

E-mail: hales@fi.muni.cz
http://nlp.fi.muni.cz/poc_lingv/

Obsah:

Co to je korpus?

- Co to je text, dokument?
  - lecos
- Různé typy korpusů
  - textové
  - mluvené
- Pro potřeby NLP
  - textový korpus

Textový korpus

- soubor textů
- charakteristiky
  - rozsáhlý (stovky mil. až mld. pozic/slov)
  - v jednotném formátu
  - stukturovaný
  - v elektronické podobě

Typy korpusů

- vždy záleží na účelu a způsobu použití
- možnosti
  - jazyk
  - typy textů
  - zdroj dat
  - značkování
  - ...

První korpus

Brown

- americká angličtina (1961)
- Brown University, 1964
- gramatické značkování, 1979
- 500 textů, 1 mil. slov
- W. N. Francis & H. Kučera
  - první statistické charakteristiky angličtiny
  - relativní četnosti slov a slovních druhů

SUSANNE

SUSANNE

- autor Geoffrey Sampson, Sussex University
- kniha *English for the Computer*
- část korpusu Brown ($\frac{1}{4}$)
- nové gramatické značkování
- syntaktické značkování

BNC

British National Corpus

- britská angličtina, 10% mluva
- první velký korpus pro lexikografy
- vydavatelé slovníků (OUP) + univerzity
- 1991–1994, World Edition 2000
- ≈3000 textů, 100 mil. slov
- gramatické značkování automatickým nástrojem

BoE

Bank of English

- britská angličtina
- COBUILD (HarperCollins), University of Birmingham
- 1991, stále rozšiřován
- 2002, ≈450 mil. slov

Další národní korpusy

- Český národní korpus
  - ÚČNK, FF UK
  - SYN2000: 100 mil. slov
  - Litera, Synek, BMK, …
- Slovenský, Maďarský, Chorvatský, …
- Americký

Korpusy na FI

vytvořené na FI, příklady:

- Desam
  - 1996, ručně značkovaný (desambiguovaný)
  - ≈1 mil. slov
- WWW
  - periodika z webu, z let 1996–1998
  - ≈100 mil.
- Chyby
  - práce studentů předmětu Základy odb. stylu s vyznačenými chybami
  - ≈400 tis.

Korpusy na FI

spolupráce

- Dopisy
- Mluv
- Kačenka
- ČNPK
- 1984
- Otto
- Italian
- Giga Chinese
- Francouzský, Slovinský, Britská angličtina, …

Formáty korpusů

- archiv/kolekce
  - různé formáty, podle zdroje/typu
- textové banky
  - jednotný formát a základní struktura
  - dokumenty/texty, základní metainformace
- vertikální text
- binární data v aplikaci
  - pomocné data pro rychlejší zpracování
    - indexy
    - statistiky

Kódování znaků

- 8 bitů $\approx$ 256 znaků
  - ASCII – základ 7 bitů
  - kódování pro češtinu
    - ISO-Latin-2, Windows-1250, 852
- Unicode
  - 32bitů na znak
  - UTF-8
    - 1 až 4 byty na znak
  - UTF-16
    - 2 až 4 byty na znak

Kódování metainformací

- escape-sekvence
  - speciální znak mění význam následujících znaků
  - `\n`, `\t`, `&`, `<tag>`
- SGML
  - Standard Generalised Markup Language
  - ISO 8879:1986(E)
- XML
  - Extensible Markup Language
  - W3C, 1998

XML

- struktura popsána v DTD
- elementy
  - počáteční, koncová značka
  - `<doc>`, `<head>`, `</head>`, `<g/>`
- atributy elementů/značek
  - `<doc title="Jak pejsek ..." author="Čapek">`
  - `<head type="main">`
- entity
  - `>`, `<`, `&`, `é`

Standardy pro ukládání

- SGML/XML
- TEI
  - Text Encoding Initiative
  - TEI Guidelines for Electronic Text Encoding and Interchange
- CES, XCES
  - Corpus Encoding Standard

Obsah korpusu

Co je v korpusu uloženo?

- text
- metainformace
- struktura dokumentu
  - odstavce, nadpisy, verše, věty
- značkování
  - informace o slovech
  - morfologie, základní tvary

Tokenizace

Rozdělení textu do pozic

- token (pozice) = základní prvek korpusu
- většinou slovo, číslo, interpunkce
  - bude-li, don't
- může silně ovlivnit výsledky

Vertikální text

- jednoduchý formát i jeho zpracování
  - každý token na samostatném řádku
  - struktury formou XML elementů
  - značkování odděleno tabulátorem
- podrobnosti
  - `http://www.fi.muni.cz/nlp/`
  - Informace pro současné a potenciální spolupracovníky
  - Textové korpusy
  - Popis vertikálů

Zpracování textů na UNIXu

- coreutils
  - cat, head, tail, wc, sort, uniq, comm
  - cut, paste join, tr
- grep
- awk
- sed / perl

Příklady použití coreutils

- slovník z vertikálního textu

```
cut -f 1 -s desam.vert |sort |uniq -c \
|sort -rn >desam.dict
```

- jednoduchá tokenizace

```
tr -cs 'a-zA-Z0-9' '\n' <GPL >GPL.vert
cat GPL.vert |sort |uniq -c |sort -rn >GPL.dict
```


Korpusové manažery

nástroje na zpracování korpusů

- uložení textu
- editace/příprava textu
- značkování
- rozdělení do pozic (tokenizace)
- vyhledávání (konkordance)
- statistiky

Systém Manatee

- korpusový manažer
- přímo podporuje
  - uložení textu
  - vyhledávání (konkordance)
  - statistiky
- externí nástroje
  - značkování
  - rozdělení do pozic

Systém Manatee

hlavní zaměření

- velké korpusy
- rozsáhlé značkování
  - morfologické, syntaktické, metainformace
- návaznost na další aplikace/nástroje
  - korpusový editor, tvorba slovníků
- univerzálnost
  - různé jazyky, kódování, systémy značek

Klíčové vlastnosti

- modulární systém
- přístup z různých rozhraní
  - grafické uživatelské rozhraní (Bonito)
  - aplikační programové rozhraní (API)
  - příkazový řádek
- rozsáhlá data
  - až 2 mld. pozic
  - neomezeně atributů a metainformací
- rychlost
  - vyhledávání, statistiky

Klíčové vlastnosti

- multihodnoty
  - zpracování víceznačných značkování
- dynamické atributy
  - vyhledávání a statistiky na počítaných datech
- subkorpusy
- silný dotazovací jazyk
  - dotazy na všechny atributy, metainformace
  - pozitivní/negativní filtry

Klíčové vlastnosti

- frekvenční distribuce
  - víceúrovňová
  - všechny atributy a metainformace
- kolokace
  - různé statistické funkce