PA153 Zpracování přirozeného jayzka

06 – Korpusy, nástroje, anotace

Karel Pala, Vít Suchomel

CZPJ, FI MU, Brno

7. listopadu 2018

1. Corpora
   - Co je korpus
   - Tradiční textové korpusy
   - Webové korpusy
   - Parallel and other corpora
2. Corpus tools
   - Nástroje k získávání korpusů
   - Korpusové manažery
   - Možnost spolupráce
3. Anotace
   - Co jsou anotace
   - Druhy
   - Vybrané otázky anotace
4. Literatura

Definice

Korpus je soubor dat (textů) v přirozeném jazyce.

Použití

- obecně: data ke studiu přirozeného jazyka
- lexikografové: slovníky
- lingvisté: jazykové analýzy, změny jazyka
- sociologové: jak a o čem píšeme, která témata jsou aktuální
- marketingoví experti: hodnocení značek a výrobků v textech
- statistické nástroje ZPJ: jazykové modely pro značkovače, analyzátory, překladové systémy, prediktivní psaní,. . .

Příklady zdrojů dat

- tištěná média: knihy, časopisy, noviny, básně
- internet: články, prezentace, blogy, diskuze, tweety
- řeč: přepis záznamů řeči, filmové titulky
- ostatní: osobní korespondence, školní eseje

Zvláštní vlastnosti korpusů

- podle data vzniku obsahu: synchronní x diachronní
- (použití diachronních: např. zkoumání trendů v používání slov)
- jednojazyčné x vícejazyčné
- paralelní x paralelní srovnatelné
- (příklad srovnatelných: stejné články Wikipedie v různých jazycích)
- podle zkrácení dokumentů: plné texty x zkrácené vzorky
- média: audio (záznam dialogu), video (záznam emocí)

Tradiční textové korpusy

Vznik

- obvykle na objednávku vládní instituce, univerzity nebo nakladatelství
- zdroje: obvykle z tištěných médií – nakladatelství, skenování knih, přepisy rozhovorů

Výhody tradičních korpusů

- kontrolovaný obsah (vyvážená reprezentace žánrů a stylů)
- kvalitní a bohaté informace o datech (autor, název, rok vydání, žánr, styl, oblast)
- možnost opravy chyb

Nevýhody tradičních korpusů

- nedostatenčná velikost pro některá použití
- obtížné získávání dat, vysoké náklady
- problémy s autorskými právy

Standard Corpus of Present-Day American English (Brown corpus)

- Brown University (Henry Kucera, W. Nelson Francis)
- 1964 (1971, 1979)
- 500 samples of text, 2000 words each = 1 million words
- `http://khnt.aksis.uib.no/icame/manuals/brown/`

British National Corpus (BNC)

- Oxford University, Longman
- 1991–1994 (2001, 2007)
- text samples, 100 million words together
- 90 % written, 10 % spoken
- `http://www.natcorp.ox.ac.uk/`

Corpus of Contemporary American English (COCA)

- Brigham Young University (Mark Davies)
- since 1990, 20 million words added each year
- 450 million words (2013)
- `http://corpus.byu.edu/coca/`

Český národní korpus SYN

- Ústav ČNK na FF UK v Praze
- texty od 1990 vydání SYN2000, SYN2005, SYN2010
- 1,3 mld. slov (2010)
- `http://korpus.cz/`

Korpus DESAM

- CZPJ FI MU
- morfologicky označkovaný korpus českých textů
- desambiguované (jednoznačné) značkování
- 1 mil. slov

1. Corpora
   - Co je korpus
   - Tradiční textové korpusy
   - Webové korpusy
   - Parallel and other corpora
2. Corpus tools
   - Nástroje k získávání korpusů
   - Korpusové manažery
   - Možnost spolupráce
3. Anotace
   - Co jsou anotace
   - Druhy
   - Vybrané otázky anotace
4. Literatura

Web je největší korpus

Myšlenka a iniciativa „Web as Corpus“ (`http://sigwac.org.uk/`)
Výhody internetových korpusů

- obrovské množství dat
- dokumenty různých druhů
- aktuální podoba psané formy jazyka
- snadná dostupnost, nízké náklady

Nevýhody internetových korpusů

- neuspořádanost
- nežádoucí obsah
- duplicity
- chyby
- víme, co stahujeme?

Proč potřebujeme velké korpusy?

Přínosy velkých korpusů

- větší slovník (více různých slov)
- více/lepší příklady použití slov ve větách
- lepší pokrytí řídkých jazykových jevů
- více dat pro přesnější jazykové modely

Velké textové korpusy získané z internetu v CZPJ

jazyk	velikost korpusu [GB]	velikost korpusu [10^9 tokenů]	doba stahování [dny]
enTenTen12	108	17.8	17
esAmTenTen11	44	8.7	14
arTenTen12	58	6.6	28
czTenTen12		5.8	40
frTenTen12	72	12.4	15
jpTenTen11	61	11.1	28
ruTenTen12	198	20.2	14
turkish web	26	4.1	14
enTenTen15		30	30

V CZPJ máme k dispozici a umíme efektivně ukládat také kolekci dat ClueWeb ‘09 — vyčištěná anglická část obsahuje zhruba 70 miliard tokenů.

1. Corpora
   - Co je korpus
   - Tradiční textové korpusy
   - Webové korpusy
   - Parallel and other corpora
2. Corpus tools
   - Nástroje k získávání korpusů
   - Korpusové manažery
   - Možnost spolupráce
3. Anotace
   - Co jsou anotace
   - Druhy
   - Vybrané otázky anotace
4. Literatura

Paralelní korpus InterCorp

- Ústav ČNK na FF UK v Praze
- jazykové páry (vždy s češtinou) zarovnané na větách
- 10–30 mil. slov každý pár
- `http://korpus.cz/intercorp/`

Další paralelní korpusy

- OPUS – všechna veřejně dostupná paralelní data (`http://opus.lingfil.uu.se/`)
- Europarl – jednání EP (`http://www.statmt.org/europarl/`)
- 1984 – Orwellův román (`http://nl.ijs.si/ME/Vault/CD/docs/1984.html`)

Google Books Ngrams

- Vyhledávání ve skenovaných knihách
- Pouze n-tice slov ($n \in \{1..5\}$)
- `https://books.google.com/ngrams`

1. Corpora
   - Co je korpus
   - Tradiční textové korpusy
   - Webové korpusy
   - Parallel and other corpora
2. Corpus tools
   - Nástroje k získávání korpusů
   - Korpusové manažery
   - Možnost spolupráce
3. Anotace
   - Co jsou anotace
   - Druhy
   - Vybrané otázky anotace
4. Literatura

Postup získávání webových korpusů v CZPJ

- příprava jayzkově závislých modelů používaných v dalších krocích — učení na dokumentech z Wikipedie
- spuštění crawleru (`SpiderLing`)
- zpracování a vyhodnocování během běhu crawleru
  - detekce znakové sady dokumentu (`Chared`)
  - filtrování jazyka (vektor trigramů znaků)
  - odstraňování nežádoucího obsahu (`Justext`)
  - kontrola duplicitních dokumentů
  - vyhodnocování průběžné výtěžnosti webových domén
- zpracování získaných dat
  - odstranění podobných odstavců (`Onion`)
  - tokenizace (`Unitok` nebo jiný nástroj)
  - značkování morfologické a syntaktické — externími nástroji, jsou-li dostupné
  - zakódování a nahrání do korpusového manažeru (`Manatee`/`Bonito`)

Web crawler

Web crawler je druh počítačového programu

- prochází internet (stránky propojené odkazy)
- stahuje dokumenty (metainformace, obsah)
- ukládá části dokumentů v různých formátech k dalšímu použití
- plánovač crawleru: určuje, které webové stránky se budou dále stahovat, např. podle vzdálenosti od počátku nebo výtěžnosti

Crawlery

- k získávání obsahu dokumentů – GoogleBot (navíc k indexování), Heritrix a mnoho dalších
- ke sbírání odkazů
- k získávání textových dokumentů pro ZPJ – SpiderLing

Ukázka dat v korpusu – XML vertikální formát

```
<dokument zanr="blog"
  nazev="Dovolená v Paříži" datum="2011-10-28"
  url="http://karel.bloguje.cz/dovolena-v-parizi">
<odstavec nadpis="1">
<veta>
Po
sedmi
letech
v
kouzelné
Paříži
!
</veta>
</odstavec>
...
</dokument>
```

Korpusový manažer

Korpusový manažer slouží k ukládání textů a práci s textovými korpusy.

- příprava textu – převod z různých formátů
- zahrnutí metadat (informací o datech – zdroj, autor, téma, žánr, datum)
- tokenizace (rozdělení na slova, interpunkce, znaky)
- anotace (značkování)
- efektivní uchování korpusu – datové struktury umožňující rychlé získání uložených dat
- konkordance – získání úseků textů odpovídajících uživatelským dotazům
- výpočet statistik – vyhledání typických vzorů v datech, frekvenční distribuce, souvýskyty

Word Sketch Engine

- korpusový manažer (a více)
- vyvíjený od roku 2000 v CZPJ FI MU (dizertační práce Pavla Rychlého)
- od 2003 spolupráce s průmyslovým partnerem Lexical Computing
- hlavní komponenty
  - Manatee – korpusový manažer
  - Bonito – uživatelské rozhraní a API
  - Corpus Architect – vytváření uživatelských korpusů a jejich nahrávání do Manatee
- pro zaměstnance a studenty MU zdarma na `https://ske.fi.muni.cz`

Manatee – korpusový manažer

- akceptuje XML vertikální formát dat
- podporuje metadata a anotace, jsou-li správně předzpracovány
- korpusy uchovává efektivně
- konkordance – získání úseků textů odpovídajících uživatelským dotazům
- Word Sketch = slovní profil – stručný přehled kolokačního a gramatického chování slova
- výpočet statistik – vyhledání typických vzorů v datech, frekvenční distribuce, souvýskyty
- *více v předmětu PA154 Statistické nástroje pro korpusy (jaro 2014)*

Corpus Query Language (CQL)

- dotazovací jazyk podporovaný Manatee
- slouží k vyhledání tokenů v korpuse
- využívá regulárních výrazů
- příklad: `[lemma="červený"|lemma="černý"] [tag="k1.*nP.*"]`
  dvě bezprostředně následující slova, první má základní tvar ,,červený“ nebo ,,černý“, druhé je podstatné jméno v množném čísle, například ,,červenými domky“ je platný odpovídající výraz

Bonito – uživatelské rozhraní a API

- převádí uživatelské dotazy do CQL
- volá funkce Manatee
- výsledek zobrazuje uživateli nebo ve formátu JSON pro API
- ukázka: `https://ske.fi.muni.cz`

Corpus Architect – uživatelské korpusy

- zajišťuje autentizaci a přístup uživatelů k jejich korpusům
- ukládá a zpracovává uživatelská data
- zpracovaná data nahrává do Manatee
- obsahuje univerzální tokenizaci
- pracuje s morfologickými analyzátory pro více než 10 jazyků
- zahrnuje nástroj WebBootCaT k získávání korpusů z internetu

Alternativy k některým funkcím Sketch Engine

- samostatné vyhledávací nástroje pro daný korpus (např. BNC)
- WordSmith (Mike Scott, `http://www.lexically.net/wordsmith`)
- AntConc (Laurence Anthony, `http://www.antlab.sci.waseda.ac.jp/antconc_index.html`)

1. Corpora
   - Co je korpus
   - Tradiční textové korpusy
   - Webové korpusy
   - Parallel and other corpora
2. Corpus tools
   - Nástroje k získávání korpusů
   - Korpusové manažery
   - Možnost spolupráce
3. Anotace
   - Co jsou anotace
   - Druhy
   - Vybrané otázky anotace
4. Literatura

Možnost spolupráce na korpusových nástrojích

- Centrum zpracování přirozeného jazyka – `https://nlp.fi.muni.cz/`
- laboratoř ZPJ – místnost B204
  - Přijďte se podívat na seminář,
  - nebo napište doc. Horákovi
- závěrečné práce
- jarní předměty
  - IB047 Úvod do korpusové lingvistiky a počítačové lexikografie
  - PA107 Projekt z korpusových nástrojů
- podzimní předměty
  - PB106 Projekt z korpusové lingvistiky
  - IA161 Pokročilé techniky zpracování přirozeného jazyka

1. Corpora
   - Co je korpus
   - Tradiční textové korpusy
   - Webové korpusy
   - Parallel and other corpora
2. Corpus tools
   - Nástroje k získávání korpusů
   - Korpusové manažery
   - Možnost spolupráce
3. Anotace
   - Co jsou anotace
   - Druhy
   - Vybrané otázky anotace
4. Literatura

Anotace

Anotace je přidávání informací (zejm. o slovech, větách nebo dokumentech) do textového korpusu. Slouží k označení a následnému zkoumání vlastností slov, vět, nebo celých textů.

- informace o zpracování dat (např. rozdělení na tokeny)
- metadata textů (zdroj, autor, téma, žánr, datum)
- struktury (dokument, odstavec, věta, zarovnání, mluvčí)
- značkování – přiřazení značky (např. slovního druhu) k tokenu

Druhy anotace

- morfologická (slovní druh a jiné gramatické kategorie)
  - u nás (čeština): morfologický analyzátor Majka
  - jiné: TreeTagger (enTenTen12), CLAWS (BNC, COCA), FreeLing (esTenTen11)
- syntaktická (parsing – závislostní nebo složkové stromy, chunking – rozdělení na fráze jmennou /NP/, slovesnou /VP/, předložkovou /PP/)
  - u nás (čeština): Synt, SET, DIS/VADIS, IOBBER (polština)
  - jiné: MST Parser, MaltParser
- sémantická (word sense tagging/desambiguation /WSD/ – rozlišení významu slova, named entity recognition – rozpoznání jmenných entit /NER/)
  - u nás (čeština): DESAMB – desambiguace morfologických značek
  - jiné: WordNet, SuperSenseTagger – WSD, NER
- koreference (určení anafory)
  - u nás (angličtina): SARA
- pragmatická (označení mluvčího, komunikační situace)

Ukázka anotací v korpusu – XML vertikální formát

```
<dokument zanr="blog" nazev="Dovolená v Paříži">
<veta nadpis="1">
Po        po        k7c6           0  8
sedmi     sedm      k4c6           1  7
letech    léto      k1gNnPc6       2  7
v         v         k7c6           3 10
kouzelné  kouzelný  k2eAgFnSc6d1   4  9
<entita druh="město">
Paříži    Paříž     k1gFnSc6       5  9
</entita>
!         !         kx             6 11
<NP>                               7  8
<PP>                               8 11
<NP>                               9 10
<PP>                              10 11
<S>                               11  -
</veta>
```

Editory anotací

- výstup vždy v XML
- GATE `http://gate.ac.uk/`
- Brat `http://brat.nlplab.org/`
- WordSmith `http://www.lexically.net/wordsmith`
- u nás: Phrase Annotator (shallow parsing: fráze, závislosti), Sysel (sémantické kategorie)

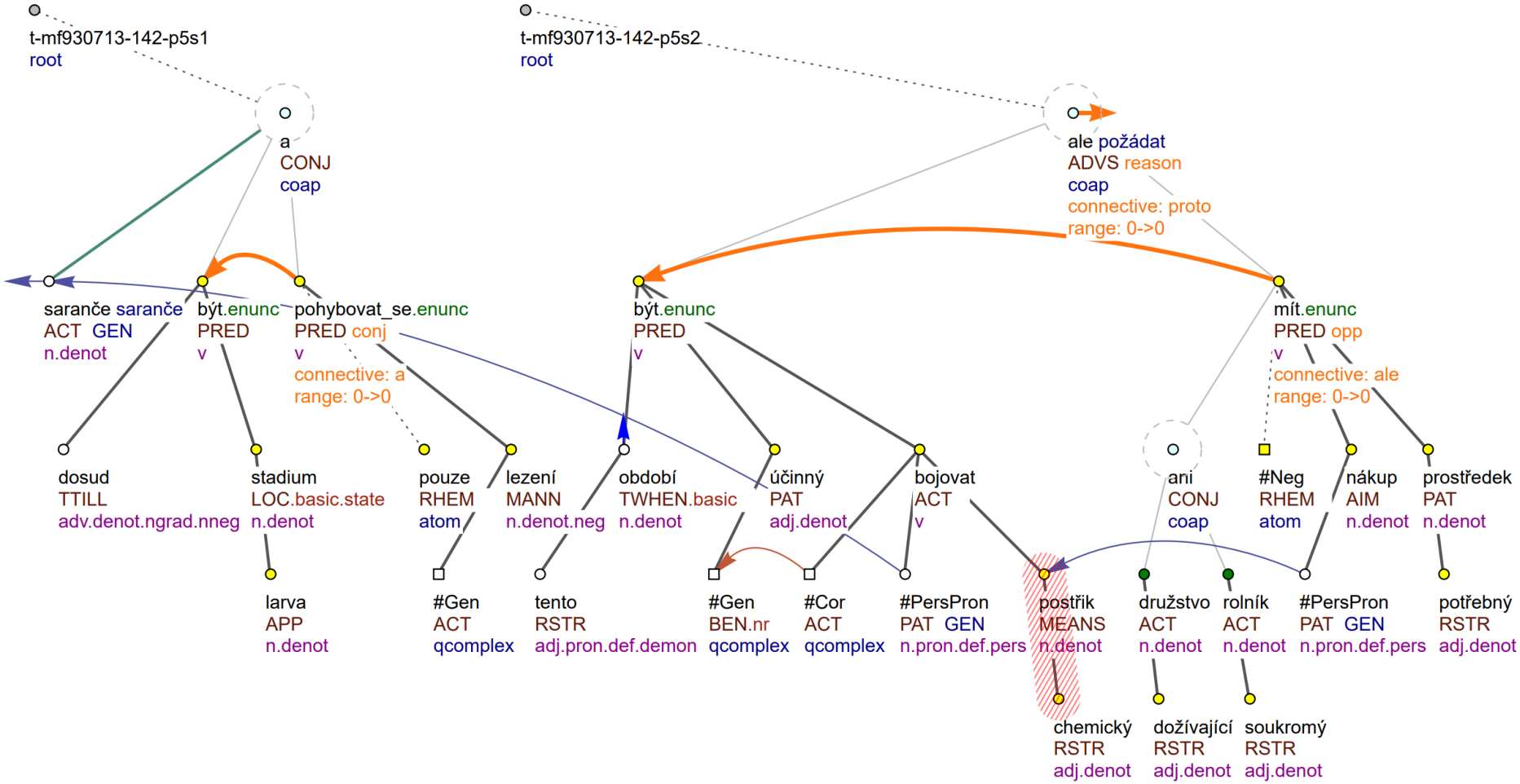
Prague Dependency Treebank

PDT je bohatě anotovaným korpusem vytvořeným v ÚFAL.
Verze 3.5 z 2018: 49 431 vět obsahujících 2 000 000 slov.
Korpus je anotovaný v rovinách

- morfologické
- syntaktické
- tektogramatické

`https://ufal.mff.cuni.cz/prague-dependency-treebank`
Hajič et al., 2018. Prague Dependency Treebank 3.5. Institute of Formal and Applied Linguistics, LINDAT/CLARIN, Charles University, LINDAT/CLARIN PID: http://hdl.handle.net/11234/1-2621.

Prague Dependency Treebank – ukázka



Tektogramatická anotace s koreferenčními odkazy (modré a hnědé šipky), víceslovné entity (rudé pruhy), anotace diskurzu (oranžové šipky a atributy): „*Saranče jsou doposud ve stadiu larev a pohybují se pouze lezením. V tomto období je účinné bojovat proti nim chemickými postřiky, ale dožívající družstva ani soukromí rolníci nemají na jejich nákup potřebné prostředky.*“

Zdroj: `https://ufal.mff.cuni.cz/pdt3.5`

1. Corpora
   - Co je korpus
   - Tradiční textové korpusy
   - Webové korpusy
   - Parallel and other corpora
2. Corpus tools
   - Nástroje k získávání korpusů
   - Korpusové manažery
   - Možnost spolupráce
3. Anotace
   - Co jsou anotace
   - Druhy
   - Vybrané otázky anotace
4. Literatura

Manuální x automatická

- Ruční anotace je zdlouhavá a nákladná. Přesto nemusí být dokonalá.
- Nedokonalá automatická anotace (naučená na ručně anotovaných datech) je pro velká data nevyhnutelná.

Anotace rozsáhlých dat

Obvyklým postupem je

1. Rozmyslet problém, vytvořit stručný anotační manuál.
2. Nechat lidi anotovat část dat.
3. Natrénovat klasifikátor problému metodami strojového učení.
4. Označkovat zbytek dat klasifikátorem.
5. Je možno prozkoumat chyby klasifikátoru, vyhodnotit mezianotátorskou shodu, poučit se z chyb anotátorů a celý proces opakovat.

Příklad: morfologická anotace korpusu.

Aktivní učení

Jaké případy vybrat anotátorům k ručná anotaci, je-li

- ruční anotace zdlouhavá $\Rightarrow$ nákladná,
- mnoho případů k dispozici $\Rightarrow$ stačí nějaký vybrat.

Příklad: určení žánru dokumentu.

Aktivní učení

Jaké případy vybrat anotátorům k ruční anotaci, je-li

- ruční anotace zdlouhavá $\Rightarrow$ nákladná,
- mnoho případů k dispozici $\Rightarrow$ stačí nějaký vybrat.

Příklad: určení žánru dokumentu.

Uncertainty sampling:

- Select the unlabelled instances that are confusing, i.e. closest to the threshold.
- Train on labelled instances, select the most uncertain unlabelled instance, label, retrain.
- The most uncertain = the least confident = the instance with the lowest certainty of the classifier.
- Classifier certainty – distribution of class predictions.

1. Corpora
   - Co je korpus
   - Tradiční textové korpusy
   - Webové korpusy
   - Parallel and other corpora
2. Corpus tools
   - Nástroje k získávání korpusů
   - Korpusové manažery
   - Možnost spolupráce
3. Anotace
   - Co jsou anotace
   - Druhy
   - Vybrané otázky anotace
4. Literatura

Literatura

- Kilgarriff, Adam, Gregory Grefenstette. Introduction to the special issue on the web as corpus. In Computational linguistics 29.3 (2003): s. 333-347.
- RYCHLÝ, Pavel a Pavel SMRŽ. Manatee, Bonito and Word Sketches for Czech. In Proceedings of the Second International Conference on Corpus Linguisitcs. Saint-Petersburg: Saint-Petersburg State University Press, 2004. s. 124-132, 9 s.
- KILGARRIFF, Adam, Pavel RYCHLÝ, Pavel SMRŽ a David TUGWELL. The Sketch Engine. In Proceedings of the Eleventh EURALEX International Congress. Lorient, France: Universite de Bretagne-Sud, 2004. s. 105-116, 12 s.
- Corpus Query Language ve Sketch Engine: `http://trac.sketchengine.co.uk/wiki/SkE/CorpusQuerying`
- Lekce Corpus Mark-up od Matthew Brook O'Donnela z UoL Summer Institute in Corpus Linguistics: `www.lexically.net/courses/sessions/markup/Corpus%20Mark-up.ppt`

Otázky, které by mohly být položeny

- Uveďte příklad dokumentů ve vícejazyčných paralelních srovnatelných korpusech.
- K čemu můžeme použít diachronní korpus?
- Srovnejte korpusy sestavené z internetových textů s korpusy z tradičních zdrojů.
- O čem rozhoduje plánovač (scheduler) v crawleru?
- Vyjmenujte alespoň pět funkcí dobrého korpusového manažeru a stručně vysvětlete, k čemu slouží.
- Jaké druhy anotace textů znáte?
- Jak byste postupovali při morfologické anotaci velkého webového korpusu?