K. Osolsobě

Výuka: od 26. 2. 2024 každých 14 dní kontaktní, každých 14 dní úkol (viz harmonogram)

Podmínky ukončení: Průběžné plnění úkolů (5 odevzdaných úkolů, závěrečný test).

Náplň dnešní hodiny 26. 2. 2022

Co je to korpus?

- Soubor textů
- elektronicky uložených a přístupných (korpusové manažery – programy, skrze něž lze ke korpusům přistupovat)
- má stanovený obsah (složený z textů záměrně vybraných dle zveřejněných kritérií)
- má stanovený rozsah/velikost (lze na něm pracovat s frekvenčními/kvantitativně měřitelnými údaji)
- obsahuje standardní anotace (metadata – údaje o každém textu a lingvistické interpretace, anotace jazykových jednotek – vnitřní anotace)

Registrace uživatele pro práci s ČNK (http://ucnk.ff.cuni.cz/)

Korpusový manažer

Základy práce s korpusem přes Kontext

KonText

Dotaz

Výběr korpusu

https://wiki.korpus.cz/doku.php/cnk:uvod

Jaké korpusy jsou k dispozici ?

Časové hledisko (synchronní / diachronní)

Hledisko textů (psané / mluvené, připravené/spontánní)

Hledisko žánru (vyvážené žánrově/ žánrově kompaktní – např. korpusy výhradně publicistické, nebo korpus soukromé korespondence, projekt Korpus českého verše).

Hledisko autora (autoři jsou rodilí mluvčí/ autoři se učí jazyk, v němž jsou texty vytvořeny jako tzv. druhý jazyk – learner corpora/žákovské korpusy, autorské korpusy založené na díle/korespondenci význačných osobností).

Hledisko jazyka (jednojazyčné – např. čeština/ vícejazyčné, srovnatelné, paralelní).

Vícejazyčné paralelní korpusy – stejné texty – originál+překlad – zarovnaní/alignment = jednotky, které si odpovídají, jsou propojeny / srovnatelné korpusy – různojazyčné i stejného jazyka vybudované stejným způsobem, mající stejné složení).

Jak čteme informace o zvoleném korpusu?

Proč je třeba citovat korpusy?

Korpusy ÚČNK vznikly jako výsledek státní podporu GAČR. V korpusech jsou texty, které mnohdy spadají pod autorská práva. ÚČNK poskytl smluvní garance „poskytovatelům textů".

Jak číst informace o velikosti korpusu:

Termíny: viz http://wiki.korpus.cz/doku.php

http://wiki.korpus.cz/doku.php/pojmy:prehled_pojmu

Tokenizace

Token je nejmenší jednotka textu, většinou se jedná o grafické slovo (tj. řetězec alfabetických znaků oddělený mezerou v textu), resp. o jednu jeho konkrétní realizaci. V některých případech je jedno grafické slovo rozděleno na dvě (např. *mohu -li*), často je také z praktických důvodů (pro snadné vyhledávání) oddělována interpunkce od předcházejícího slova (3 tokeny: *řekl , že*). O jednotlivých tokenech v korpusu se také mluví jako o pozicích.

Všimněme si:

Pro rodilého mluvčího je zvratné *se* součástí reflexiva tantum *ptát se*. Pro automatickou morfologickou analýzu jde o dva samostatné tokeny.

Pozice

V souvislosti s tím, že každý text, který vstupuje do korpusu, prochází procesem tokenizace, se o jednotkách v korpusu nemluví jako o slovech, ale častěji jako o **pozicích**. Tokenizace se přitom u jednotlivých korpusů může lišit, pozicí se tak v různých korpusech může myslet různě vymezená jednotka.

Lemmatizace a taggování

Lemma je reprezentativní slovníková podoba hesla, při automatickém zpracování jazyka je pak tato podoba v procesu lemmatizace přidělována každé formě v korpusu.

Přístupy k lemmatizaci se mohou v drobnostech lišit, obecně však platí, že

- lemma každého českého substantiva je jeho **nom. sg.** (tvary *lesům, lesy, lesích* mají lemma ***les***)
- u adjektiv je to **nom. sg. mask.pozitiv** (tvary *chytrého, chytrou, chytrejma, nejchytřejší* mají lemma ***chytrý***)
- u zájmen je to **nom. sg. mask.** (tvary *ta, to, ti, tomu* mají lemma ***ten***)
- u sloves je to **infinitiv** (tvary *chodil, chodíš, chodíme* mají lemma ***chodit***)

Lemma jako jednotka vzniká abstrakcí morfologických vlastností slovního tvaru (označovaného jako word nebo forma), představuje tedy množinu forem se stejným kořenem lišící se pouze morfologickými afixy, příp. pravopisnou variantou. V některých koncepcích se pak k lemmatu řadí i vybrané varianty slovotvorné.

Představme si následující dialog, která z variant je podle vás více na místě, A nebo B?

A

- No víš, viděl jsem takovou ***fuškunkuli*** a ona ti měla na hlavě takovou ***kumušinku paškovanou*** a ona si ji ještě ***vygárovala***.

- Co je to ***fuškunkuli*** a ***kumušinku paškovanou***? A co je to ***vygárovala***?

B

- No víš, viděl jsem takovou ***fuškunkuli*** a ona ti měla na hlavě takovou ***kumušinku paškovanou*** a ona si ji ještě ***vygárovala***.

- Co je to ***fuškunkule*** a ***kumušinkule paškovaná***? A co je to ***vygárovat***?

Uvědomte si, že lemmatizace je činnost, kterou dnes automaticky provádí řada nástrojů od vyhledávačů na webu přes on-line slovníky. Jde ale o schopnost, kterou nabývá i dítě během akvizice jazyka, kterou má mluvčí, když se dotazuje na neznámé slovo, kterou aplikujeme, když hledáme v cizojazyčném slovníku (např. význam tvaru *went* nenajdeme ve slovníku angličtiny pod ***w***, ale pod ***g***).

Desambiguace

Desambiguace (někdy též disambiguace, z lat. *dis-* vyjadřuje zápor, *ambo* oba, česky zjednoznačnění) je část (většinou automatického) procesu anotace jazykových dat, které vstupují do korpusu.

Zjednoznačněním se většinou myslí odstranění homonymie, čili jednoznačná interpretace slovního tvaru či skupiny slovních tvarů nebo věty na základě kontextu či mimojazykové situace. Desambiguace se obecně týká všech jazykových rovin, nejčastěji se ovšem v

korpusech češtiny uplatňuje na rovině morfologické (zahrnující lemmatizaci a přiřazení náležitých morfologických údajů slovnímu tvaru na základě kontextu).

Např. ve větě *Větry vanou od západu.* se při morfologické interpretaci věty nejprve přiřadí morfologickou analýzou tvaru *vanou* dvě lemmata a dvě morfologické interpretace:

1. lemma = *vana*, subst. fem. sg. instr.
2. lemma = *vát*, 3. os. pl. préz,

a poté se při desambiguaci vybere náležitá 2. interpretace.

V následujících větách si všimněte, jak je třeba nejednoznačný tvar ***sil***, který lze interpretovat jako a) genitiv plurálu feminina k lemmatu ***síla***, b) genitiv plurálu neutra k lemmatu ***silo***, c) variantní tvar l-ového příčestí maskulina singuláru slovesa ***sít***.

*a) Podle jeho názorů je internet jednou ze **sil**, která dostala Ameriku na špici*

*b) Z jednoho ze **sil** začala náhle tryskat čpící tekutina a ocelová konstrukce jedné z věží se zhroutila.*

*c) Raná variační fantazie na lidový nápěv **Sil** jsem proso dala oběma protagonistům možnost ukázat jejich virtuozitu.*

Někdy může být situace dosti složitá:

*Odstupující ministr informatiky Vladimír Mlynář podle serveru iDNES odmítl nabídku premiéra Grosse **stát se** šéfem Českého telekomunikačního úřadu.*

*Potřeboval **stát se** svým zločineckým gangem.*

Jaké přednosti má lemmatizovaný a morfologicky označkovaný korpus?

Možnosti vyhledávání v korpusu:

Nabídka výchozího atributu je závislá na konkrétním korpusu, na použité lemmatizaci a značkování.

Regulární výrazy (http://wiki.korpus.cz/doku.php/pojmy:regularni_vyrazy)

Konkordance, KWIC

Konkordance představuje všechny doklady (výskyty) hledaného jevu v korpusu spolu s **okolním kontextem**. V praxi se v rámci konkordance rozlišuje **KWIC** (tj. key word in context), tedy hledané slovo/jev a jeho pravý a levý kontext. Jeden řádek konkordančního seznamu se označuje jako konkordanční řádek.

Zobrazení

KWIC/Věta

Korpusová nastavení

(Lemma, POS – part of speech)

Metainformace

Kompletní info o zdrojovém textu: