Základy využití korpusu v praxi **cjbb75**

**Středa: 10.00-12.00 G13**

27. 2. Dú: získat přístupová práva (**uživatelské jméno/heslo**), přečíst si charakteristiky dostupných korpusů ÚČNK, znát odpovědi na otázky v hdt.

6. 3. Kontext – **příkazový řádek** – vyhledávání podle atributů (***word, lemma, tag, lc***), **cql**, **regulární výrazy**, **KWIC,** **konkordanční seznam a frekvenční seznam** jako zdroj poznání o jazykových vlastnostech vyhledaných jednotek (způsob psaní/pravopis: přejatá slova /*mailovat/mejlovat/majlovat*/, aplikace pravidel – varianty /ismus/izmus/, varianty pravopisu zdrobnělin vlastních jmen – Terinka/Terynka).

[https://wiki.korpus.cz/doku.php/manualy:kontext:novy\_dotaz](https://wiki.korpus.cz/doku.php/manualy%3Akontext%3Anovy_dotaz)



**Cvičení**

Jak česky píšeme?

**mail**

**? mailovat, meilovat, majlovat, mejlovat ?**





Složitější formulace dotazu

**[lemma="m(ai|ei|aj|ej)l.\*"]**

[**https://wiki.korpus.cz/doku.php/manualy:kontext:frekvencni\_distribuce**](https://wiki.korpus.cz/doku.php/manualy%3Akontext%3Afrekvencni_distribuce)





Můžeme sledovat frekvence pravopisných variant. Pracujeme s korpusem psaných textů, které navíc prošly jazykovou kontrolou. Úzus, který není jednotný, se může stát zdrojem pro úvahy o kodifikaci, která má odrážet úzus.

**Dvojice typu *Terinka/Terynka***

*lemma*

ter[iy]nka



Není rozpoznáno (X), tvar na –ynka není doložen.

Je opravdu takový pravopis zdrobnělin vlastních jmen možný?

Položme obecněji zadaný dotaz. Všechna podobná slova – užití regulárních výrazů.

Dotaz: Vyhledej všechna lemmata taková, že pro ně platí, že začínají libovolným znakem, za nímž následuje libovolný počet opakování libovolného znaku a končí na řetězec *inka* nebo *ynka*.



.\* [iy]nka







Přestože *Barinka* není hypokoristikum od Bára, může nám nalezení tohoto dokladu pomoci. Jak? Omezíme se na případy podobné, tedy na lemmata u nichž před *[iy]* předchází tzv. tvrdá souhláska *[rkh]*



.\*[rkh][iy]nka



To už vypadá lépe. Ale najdeme vůbec nějaké relevantní doklady?







Co nám říká IJP a co lze nalézt v korpusech







P-filtr

.\*ynka



Závěr: Je patrné, že analýza korpusových dat může přinést některé poznatky odporující běžným poučkám a pravidlům. Analýza deminutivního sufixu –*inka* užívaného pro tvoření podob vlastních jmen prokázala rozkolísanost úzu, který je v rozporu s pravidly uvedenými v IJP.

**Naučili jsme se:**

Používat některé **regulární výrazy** v jazyce ***cql***. (.\*, [], |)

Používat **statistiky** a **zobrazení.**

Používat **filtry**.

Ukázali jsme, **jak lze používat jazykový korpus k hledání odpovědí na otázky týkající se grafické podoby jazyka**.

Doplňující četba:

OSOLSOBĚ, Klára a Jana MACHALOVÁ. Hypokoristika z rodných jmen v Korpusu soukromé korespondence. In Hladká, Zdeňka a kol.. *Soukromá korespondence jako lingvistický pramen*. Vyd. 1. Brno: Masarykova univerzita, 2013. s. 33-59.

**13. 3. Dú:**

**1. Jaká jsou lemmata/tvary adaptovaných slov a jaké je frekvenční rozložení variantních způsobů jejich grafické realizace (slovesa utvořená od anglických základů: použí(va)t *Skype= ????,* použí(va)t *google = ???,* použí(va)t *twitter = ???*).**

**Uvědomte si, že nově přejatá slova nebudou ve slovníku automatické analýzy, a je tedy třeba na ní moc nespoléhat.**

**2. V IJP se uvádí, že:**

**Časté chyby – výrazy nesprávně považované za spřežky**

Jelikož se v následujících případech velmi často chybuje, připomínáme, že jen zvlášť píšeme zpodstatnělá přídavná jména, která se užívají pouze či převážně s předložkou *na* a pojí se se 4. p. Jsou to např. spojení: *na shledanou, na viděnou, na slyšenou, na rozloučenou, na odchodnou,* *na uvítanou, na zotavenou, na posilněnou, na pováženou, na uváženou, na rozmyšlenou, na vysvětlenou, na srozuměnou, na upřesněnou, na vybranou, (jako) na zavolanou*.

Dalším slovním spojením, které je často mylně chápáno, a proto psáno jako spřažený (jednoslovný) výraz, je *na rozdíl*. Také v tomto případě je nutné psát uvedené spojení jako slova dvě, respektive sekundární předložku *na rozdíl od* jako slova tři.

**Zjistěte, jaký je stav dodržování uvedené pravopisné normy v korpusech řady SYN a popište, jak jsou nesprávně (dohromady) zapsané doklady interpretovány na rovině lemmatizace a tagování.**