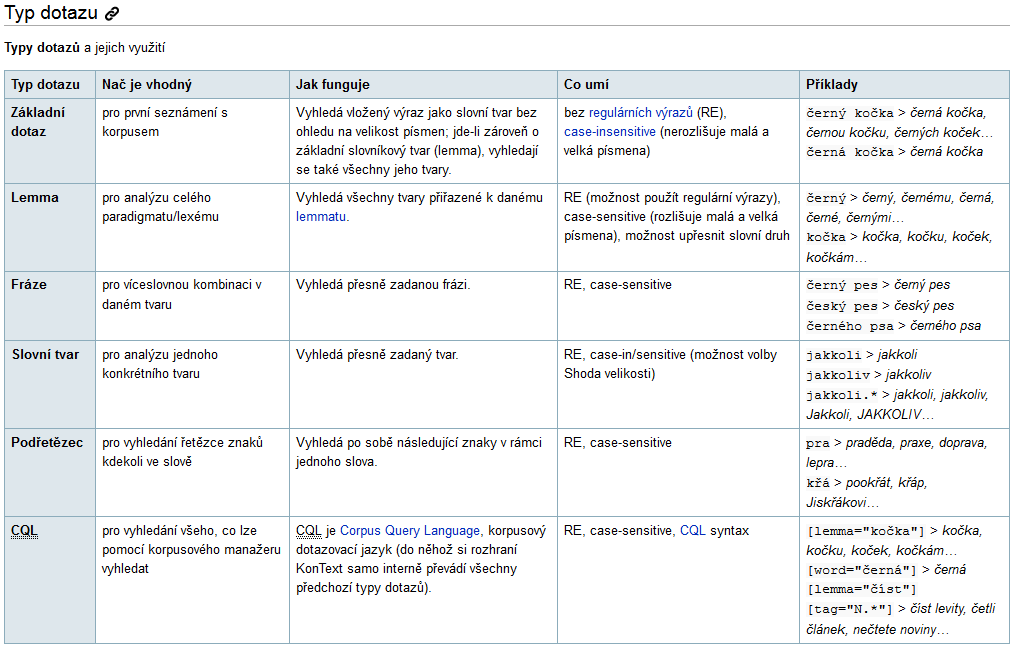
Základy využití korpusu v praxi **cjbb75\_ LS\_2022**

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>



**Cvičení**

Umíme česky?

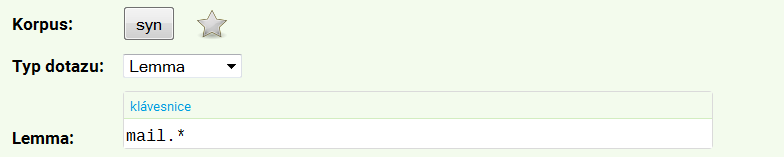
Sloveso péci je odvozovacím základem řady příbuzných slov. Při tvoření slov dochází v češtině k hláskovým alternacím. Lexikální morf mívá řadu alomorfů. Doplňte následující tabulku o všechny možné doklady příbuzných slov. Pokud si nebudete moci vzpomenout, zda je některá kombinace přípustná, vezměte si na pomoc korpus.

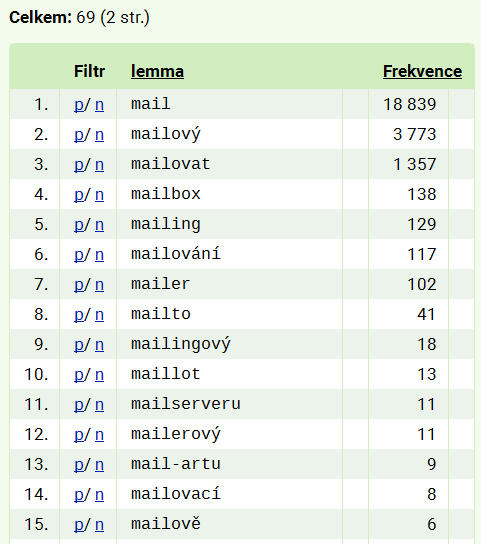
|  |  |  |  |
| --- | --- | --- | --- |
|  | P-K | P-C | P-Č |
| é | pék | péc-t | péč |
| í | pík | píc | píč |
| e | pek | pec | peč |

Jak česky píšeme?

**mail**

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

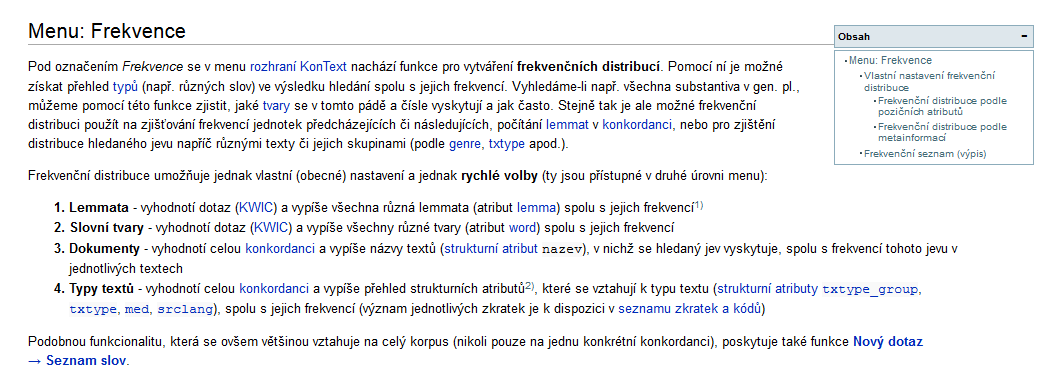


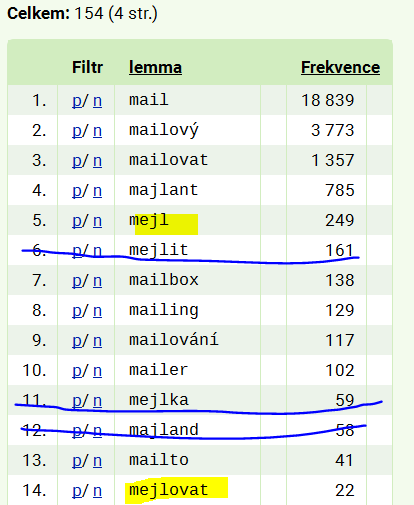


Složitější formulace dotazu

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

[**https://wiki.korpus.cz/doku.php/manualy:kontext:frekvencni\_distribuce**](https://wiki.korpus.cz/doku.php/manualy:kontext:frekvencni_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.

Úkol:

Jak se v jazykových korpusech dodržuje norma?

Obsah obrázku text

Popis byl vytvořen automaticky

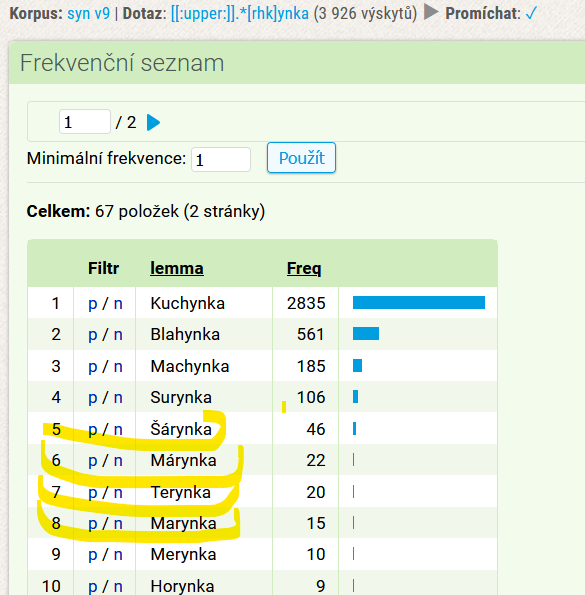
**Jak shromáždíme relevantní data pro corpusově založený výzkum?**

1. Spolehneme se na dodržení normy a vyhledáme lemmata, která splňují dvě podmínky, a sice začínají velkým písmenem a končí na řetězec inka před nímmž předchází souhlásky, po nichž se v češtině obvykle nepíše *i*. [lemma="[[:upper:]].\*[rhk]inka"]
2. Zkontrolujeme, zda se norma vždy dodržuje. Vyhledáme lemmata, která splňují dvě podmínky, a sice začínají velkým písmenem a končí na řetězec ynka před nímž předchází souhlásky, po nichž se v češtině obvykle nepíše *i*. [lemma="[[:upper:]].\*[rhk]ynka"]
3. Pozorujeme odchylky od normy. Vyhledáme lemmata, která splňují dvě podmínky, a sice začínají velkým písmenem a končí na řetězec inka nebo ynka před nímž předchází souhlásky, po nichž se v češtině obvykle nepíše *i*. [lemma="[[:upper:]].\*[rhk][iy]nka"]

Obrázek k dotazu 1



Obrázek k dotazu 2



Obrázek k dotazu 3

Obsah obrázku stůl

Popis byl vytvořen automaticky

Podívejme se na případ, kdy frekvence variant je velmi blízká a není zanedbatelně nízká, abychom ji mohli pokládat za chybu.

Šár[iy]nka



Obsah obrázku stůl

Popis byl vytvořen automaticky

Všimněme si, že v případě porušení pravopisné normy jde o proprium užité jako název instituce.

Obsah obrázku stůl

Popis byl vytvořen automaticky

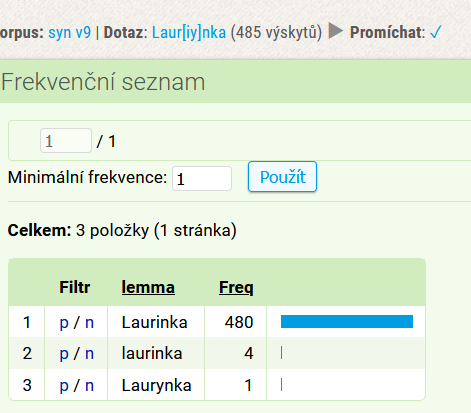
Můžeme dále zkoumat, které z frekventovaných podob deminutiv nedodržují předepsanou pravopisnou normu:

M[áa]r[iy]nka



Klár[iy]nka







Obsah obrázku stůl

Popis byl vytvořen automaticky

Obsah obrázku text

Popis byl vytvořen automaticky

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 případy úzu, který je v rozporu s pravidly uvedenými v IJP. V případě frekventovanějších užití se ovšem ukazuje, že nejde vždy o chybně utvořené deminutivum. Je třeba brát v potaz např. příjmení (Terynka), kdy hrají roli mimojazykové faktory.

Pravopisná norma vychází z existence deminutivní přípony -inka, která nemá dvě pravopisné realizace. Jestliže ovšem připustíme dvě pravopisné realizace přípony -i, kterou se v dnešním úzu frekventovaně tvoří domácké podoby proprií, pak je třeba mít na zřeteli, že tlak systému/analogie bude působit i na pravopisnou normu deminutiv na -inka.

Obsah obrázku text

Popis byl vytvořen automaticky

Obsah obrázku text

Popis byl vytvořen automaticky

**Naučili jsme se:**

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

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

Používat **Frekvence**.

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

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.

**Dú:**

7. 3. 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 = ???*). V korpusu SYNv9  vyhledejte tvary sloves tvořených od substantiva instagram a sledujte, ke kterým slovesným třídám patří.

<https://is.muni.cz/auth/el/phil/jaro2022/CJBB75/ode/ukoly_cjbb75/1_ukol/>