

Základy využití korpusu v praxi **cjbb75**

Středa: 10.00-11.30 G13

3. 4. 2019

Korpus jako zdroj dat pro výzkum slovtvorby

<http://www.ujc.cas.cz/elektronicke-slovniky-a-zdroje/Slovník-afixu-html>

<http://ufal.mff.cuni.cz/derinet>

<https://nlp.fi.muni.cz/projects/derivancze/>

Jak zadat dotaz pro hledání slov, která mají společné slovtvorné vlastnosti

(lingvistické termíny: morfém, afix, kořen, kmen, slovní druh, substantivum, adjektivum, sloveso, ..., gramatická kategorie, rod, číslo, pád, osoba ...; termíny pro formální zadání: řetězec, token, lemma, morfologická značka/tag, pozice/atribut, hodnota, technické řešení, ...)

Adjektiva od sloves typu *koupací*

Slovní formulace – tvrzení, která musí platit, podmínka nutná/dostačující, využití regulárních výrazů a logických operací, využití filtrů, pozorování dat a úprava dotazu.

Chceme vyhledat všechna adjektiva taková, že lemma končí na *cí*.

The screenshot shows a search interface titled "Hledat v korpusu". The search criteria are as follows:

- Korpus:** syn2015
- Typ dotazu:** Lemma
- Dotaz:** .*cí
- Slovní druh:** Nespecifikováno

There are two expandable sections: "Specifikovat kontext" and "Omezit hledání", both currently collapsed. A "Hledat" button is located at the bottom left.

Korpus: syn2015 | Dotaz: *ci (480 720 výskytů) ▶ Promíchat: ✓

Výskytů: 480 720 | i.p.m. 3 981,16 (vztaženo k celému "syn2015") | ARF 257 539,28 | Výsledek je promíchan 1 / 12 018

Výběr řádků: základní | Atributy:

<input type="checkbox"/>	<input type="checkbox"/>	příští sezoně v Pardubicích . Devětadvacetiletý bek se s vedením	úřadujícího	extraligového mistra dohodl na nové smlouvě . BĚL
<input type="checkbox"/>	<input type="checkbox"/>	odpady určitá rizika . Při elektronických aukcích dochází v průběhu	zadávácho	řízení (při sledování na portálu) k poklesu cen
<input type="checkbox"/>	<input type="checkbox"/>	Mussik píše : , Spousta aut , kočárů , rychle	jezdících	elektrických vozů vjžděla bez rozmyslu do hemžicíc
<input type="checkbox"/>	<input type="checkbox"/>	Bank ke včerejšímu dni celkem 54 tisíc klientů . O	zbývajících	36 tisíc zákazníků tedy může UniCredit dále usilovat
<input type="checkbox"/>	<input type="checkbox"/>	něj vede schodiště z 19 . století Kololoď a její	cestující	Cesta vzhůru Salon na horní palubě Nemůžete zablc
<input type="checkbox"/>	<input type="checkbox"/>	se opakovaná upozornění na osobní odpovědnost za jejich obsah ;	rozmožovací	přístroj nepřicházel v úvahu . Tyto obecné směrnice
<input type="checkbox"/>	<input type="checkbox"/>	zabil ji šok . Nechal ji ležet , stále s	hoříem	oděvem , obklopenou pachem seškvareného masa .
<input type="checkbox"/>	<input type="checkbox"/>	kapitalismu v Rakousku , ale přesto zůstal vášnivým idealistou ,	věřícím	v pokrok , který spatřoval především v emancipaci a
<input type="checkbox"/>	<input type="checkbox"/>	obývák zaplaví zlaté světlo . Mrkám , abych dostatečně zvlhčila	pálící	oči . Zornice se stáhnou , přesně , jak se
<input type="checkbox"/>	<input type="checkbox"/>	a bourací práce * Blok ii . Práce vnitřní a	dokončovací	* Blok iii . Vady na tepelněizolační obálce budovy ka
<input type="checkbox"/>	<input type="checkbox"/>	vzdělávání je zapotřebí v případě , že jsou splněny dvě	následující	podmínky (Buckley – Caple , 2004 , s. 32
<input type="checkbox"/>	<input type="checkbox"/>	energie : energie elektromagnetických vln a tepelné energie . Při	přepínacích	přechodných dějích vždy dochází k vyzařování elekt
<input type="checkbox"/>	<input type="checkbox"/>	dalším roce hodně dojíly . Zbytky ze středověčerního stolu putovaly	domácím	zvířatům , podarovány byly i stromy v zahradách , ab
<input type="checkbox"/>	<input type="checkbox"/>	kdy v Terry m opět zahlédl muže , který sedával u	hracích	stolu v Keithově bytě a po každém kole hbitě a
<input type="checkbox"/>	<input type="checkbox"/>	touží evropští muži ? Společnost , která jako první představí	fungující	přípravek proti vypadávání vlasů , udělá neskutečný

výsledek je značně přegenerovaný.

Lze dotaz optimalizovat?

Všimněte si:

	Filter	lemma	Freq	
1	p / n	domácí	24910	
2	p / n	pomocí	15350	
3	p / n	následující	13689	
4	p / n	vedoucí	11597	
5	p / n	budoucí	8049	
6	p / n	vynikající	5245	
7	p / n	stávající	5179	
8	p / n	rostoucí	4725	
9	p / n	cestující	4608	
10	p / n	odpovídající	4321	
11	p / n	vzdělávací	4171	
12	p / n	rozhodující	4010	
13	p / n	žádoucí	3592	
14	p / n	týkající	3445	
15	p / n	žijící	3424	
16	p / n	související	2993	
17	p / n	obývací	2969	
18	p / n	věřící	2719	
19	p / n	stojící	2600	
20	p / n	ležící	2543	
21	p / n	zbývající	2530	
22	p / n	psací	2478	
23	p / n	existující	2475	
24	p / n	pracující	2248	

obývat – obývací i obývající

zbývat – zbývající

Existují dva významy

a) vlastnost vyjadřuje, že něco je určeno k tomu, co označuje základové sloveso: *obývací kuchyň* je nejen k vaření a jedení, ale i k tomu, abychom v ní *pobývali/přebývali*;

b) transpozice významu děje (vlastnost toho, který dělá to, co označuje základové sloveso): zvíře *obývající* oblast/světadíl/ ... = jde pouze o nominalizaci verba (převod/transpozici z kategorie slovního druhu).

Rozlišujeme dvě slovotvorné třídy:

-a-cí od kmene minulého/infinitivního (tzv. **adjektiva účelová**)

-aj-í-cí od kmene přítomného (tzv. **adjektiva procesuální**)

Jak formulovat dotaz?

Pokusíme se odstranit ta adjektiva, u nichž před sufikem *-cí* předchází kmenotvorný sufix přítomného kmene. Opěrným tvarem je tvar 3. osoby sg. indikativu přítomného aktiva, tedy buď *-ou-*, nebo *-í-* (před *-í-* může ještě předcházet *-j-/-uj-/-[eě]j-/-aj-*).

Co může předcházet před *-cí* při derivaci od kmene minulého? Mohou nastat případy, kdy je kmen minulý a přítomný homonymní, takže při homonymii sufíxu budeme mít možnou homonymii adjektivních derivátů?

Existují v češtině slovesa, jejichž kmen minulý (opěrný tvar je l-ové přičestí) končí na *oul/il*?

Dokážete si vzpomenout?

Podívejte se do korpusu.

Dotaz je formulován slovně tak, že **chceme vyhledat všechny tvary slovesa l-ového přičestí maskulina singuláru takové, že končí na *oul* nebo na *íl***. V jazyce cql vypadá dotaz takto:

```
[lc=".*(oul|íl)" & tag="Vp[MI]S.*"]
```

Výsledek je prázdný konkordanční seznam.

Můžeme tudíž použít negativní filtr a odstranit lemmata na *.*(ou|í)cí*.

Sledujeme-li frekvenčně utříděný seznam, zjistíme, že se v něm vyskytují i adjektiva, která sice na *cí* končí, ale nejsou tvořena od sloves, a tak nespádají ani do kategorie účelových, ani do kategorie procesuálních adjektiv.

První je adjektivum *domáci* se vztahem k adverbium *doma*.

Zakončení je *áci*.

Můžeme si položit otázku, zda existují v češtině adjektiva od sloves, která končí na *áci*?

Můžeme si na ně a) zkusit vzpomenout, b) vyhledat pomocí pozitivního filtru všechna adjektiva na *.*áci* a prohlédnout si je, c) zkontrolovat, zda existují slovesné tvary l-ového

příčestí v singuláru maskulina na **ál* a pokud ano, tak jak se od takových tvarů tvoří adjektiva s významem účelu.

Odpověď na a): ?;

b) nalezneme pouze lemma *domáci* a kompozita se druhým členem *domáci*;

c) `[lc="*ál" & tag="Vp[MI]S.*"]` vyhledáme 126 lemmat, přičemž vidíme, že *á* se při tvoření účelového adjektiva buď krátí (*hrát* → *hrací*, *přát* → *přací*, *sát* → *sací*,), nebo se střídá s *oj* (*stát* → *stojací*);

Dále vidíme adjektiva jako *zvířecí* a *kuřecí*, jde o jinou slovotvornou třídu (adjektiv tvořených ze substantiv označující vztah přivlastnění/podobnosti základovému substantivu), ale jde také o jiný slovotvorný typ, a sice o skupinu adjektiv tvořených příponou *-í/* konverzí ke vzoru *jarní*, přičemž základem je o skupiny substantiv skloňovaných podle vzoru *kuře* kmen rozšířený o kmenotvorné *-[eě]t-*, které alternuje s *-[eě]c-*. V případě, že se základové substantivum nesklonuje podle vzoru *kuře*, pak se příslušné adjektivum tvoří pouze příponou *-í/konverzí* ke vzoru *jarní* (*psí*, *vlčí*, *mroží*, *kozí*, *myší*, ...).

Podívejme se, zda tato adjektiva mohou mít homonymní zakončení s adjektivy tvořenými ze sloves.

Vyhledáme pomocí pozitivního filtru adjektiva, která končí na *[eě]cí*.

Vidíme, že stejné zakončení mají i adjektiva jako *obráběcí*, *prováděcí*, *předváděcí*, *secí*, *sklízecí*, *dobíjecí*, *secí*, ...

Aplikovat operaci "filtr"

Filtr: negativní

Rozsah hledání: od: 0 do: 0, včetně KWIC

Typ dotazu: Lemma

Dotaz: `.*(ou|í)cí`

V dotazu CQL můžete vložit další řádek stisknutím klávesy Shift+ENTER (další tip)

Slovní druh: Nespecifikováno

Hledat

Celkem: 2 772 položek (56 stránek)

	Filter	lemma	Freq	
1	p / n	domácí	24910	
2	p / n	pomocí	15350	
3	p / n	vzdělávací	4171	
4	p / n	obývací	2969	
5	p / n	psací	2478	
6	p / n	řídící	2055	
7	p / n	zvířecí	1631	
8	p / n	parkovací	1481	
9	p / n	dýchací	1417	
10	p / n	hrací	1291	
11	p / n	ovládací	1226	
12	p / n	kuřecí	1187	
13	p / n	přijímací	1129	
14	p / n	testovací	926	
15	p / n	poznávací	883	
16	p / n	krycí	847	
17	p / n	napájecí	843	
18	p / n	měřicí	826	
19	p / n	vyšetřovací	811	
20	p / n	chladičí	797	

Jak odstranit přegenerované doklady typu *domácí, zvířecí, kuřecí, ...?*

Ruční analýza dat.

Další nástroje pro hledání slovtvorných zákonitostí.

Vyhledávání dvojic slov podle zadání společné a odlišné části pomocí nástroje Morfio.

Morfio

<- společný odlišný Morf. specifikace: +>

vzor 1:

vzor 2:

Další vzor

Korpus: Frekvence vyšší než: Hledají se: Vyhodnocují se:

Velikost písmen: ignorovat

Odkaz na toto zadání: <http://morfio.korpus.cz/nfXUOR2U>

Souhrn		Výpis		Produktivita		vzor 1		vzor	
vzor 1 ▲▼ (fq ▲▼)		vzor 2 ▲▼ (fq ▲▼)							
1	adresovací (10)	adresovat (350)							
2	aportovací (15)	aportovat (40)							
3	aranžovací (22)	aranžovat (113)							
4	aretovací (2)	aretovat (14)							
5	armovací (54)	armovat (9)							
6	asfaltovací (2)	asfaltovat (7)							
7	auditovací (4)	auditovat (21)							
8	bagrovací (2)	bagrovat (12)							
9	balíčkovací (1)	balíčkovat (4)							
10	balicí (183)	balit (1200)							
11	balzamovací (9)	balzamovat (13)							
12	barvicí (42)	barvit (478)							
13	batolecí (84)	batolit (61)							
14	bdící (62)	bdít (579)							
15	bednicí (52)	bednit (5)							
16	běhací (32)	běhat (3248)							
17	bělicí (59)	bělet bělit bílit (58)							
18	betonovací (2)	betonovat (70)							
19	biflovací (2)	biflovat (31)							
20	biřmovací (6)	biřmovat (13)							
21	bivakovací (3)	bivakovat (29)							
22	blednoucí (61)	blednout (406)							
23	blikací (3)	blikat (484)							
24	blinkací (1)	blinkat (21)							
25	blogovací (23)	blogovat (49)							
26	blokovací (39)	blokovat (959)							
27	blyštící (40)	blyštět (113)							
28	bobtnací (2)	bobtnat (137)							
29	bodací (1)	bodat (379)							

Jak vypadají výsledky korpusových analýz v lexikografickém díle zaměřeném na popis slootovorných prostředků:

<http://www.slovníkafixu.cz/heslar/-c%C3%AD>

10. 4. Dú: Jak lze v korpusu vyhledat substantiva tvořená příponou –č, která označují osoby vykonávající činnost označenou základovým slovesem: *řidič, nosič, sběrač, očišťovač, hráč*, ...

Pokuste se:

- formulovat dotaz (v přirozeném jazyce a v jazyce cql);
- pozorovat korpusová data vyhledaná na základě dotazu a dotaz specifikovat;
- popsat, jaká pozorování lze využít při reformulaci dotazu/používání filtrů;
- popsat možnosti využití nástroje Morfio k vyhledání dvojic sloveso/substantivum označující člověka, který dělá to, co označuje základové sloveso typ *nosit/nosič*;
- podívejte se na heslo –č ve Slovníku afixů užívaných v češtině a porovnejte svá pozorování s tím, co se uvádí v hesle. Narazili jste v korpusových datech na něco, co by bylo v rozporu s tím, co se uvádí v textu hesla?
- Jak byste vysvětlili slova jako *držič, sedič, tiskač, snoubič, strojič*?