

CJBB75 – 1 (G13 8.00-9.30)

K. Osolsobě

Výuka: od 2. 3. každých 14 dní kontaktní, každých 14 dní úkol (viz harmonogram)


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

Náplň dnešní hodiny

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/>)

Kdo jsme?  ČESKÝ NÁRODNÍ KORPUS

Český národní korpus je **akademický projekt** založený v roce 1994 při **FF UK** a spravovaný **Ústavem Českého národního korpusu**. Jeho cílem je systematicky mapovat češtinu a další jazyky ve srovnání s ní. **Korpusy ČNK** jsou po **bezplatné registraci** otevřeny všem zájemcům o jazyk, kteří touží vědět, jak se čeština používá. [více...](#)

Korpusový manažer

Základy práce s korpusem přes Kontext

KonText

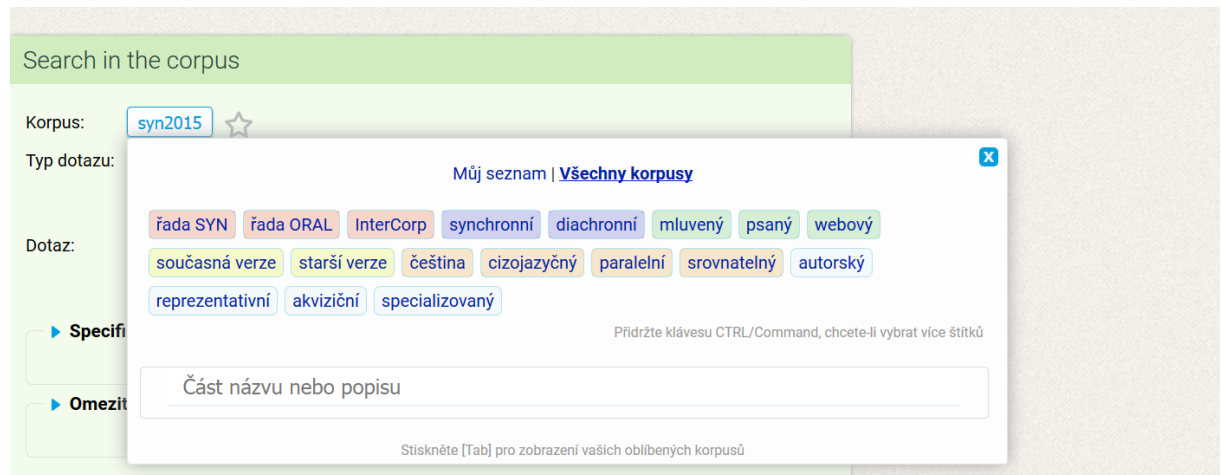


KonText Park SyD Morfio KWords Wiki Podpora Biblio Přihlášení Registrace English

ČESKÝ NÁRODNÍ KORPUS

Dotaz

Výběr korpusu



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/žakovské korpusy, autorské korpusy založené na díle/korespondenci významných osobností).

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

Vícejazyčné paralelní korpusy – stejné texty – originál+překlad – zarovnání/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?

Dotaz Korpusy Uložiti Konkordance Filtr Frekvence Kolokace Zobrazení Nápověda

Korpus: syn2015

Search in the corpus

Korpus: syn2015

Typ dotazu: Základní

Dotaz:

V dotazu, OOL můžete vložit další kód atakem klávesou Shift+ENTER

Specifikovat kontext

Omezit hledání

Hledat

syn2015

Popis: Synchronní reprezentativní korpus

Atributy: 120 748 715 pozic

Wikipedie stránka: <http://wiki.korpus.cz/doku.php/cnk:syn2015>

Dostupná metadata:

| Atributy | | Struktury | |
|----------|-----------|-----------|-----------|
| word | 1 751 599 | <doc> | 3 376 |
| lo | 1 420 255 | <text> | 114 492 |
| lemma | 777 011 | <w> | 2 905 065 |
| lemma_lc | 733 708 | <e> | 6 004 732 |
| tag | 3 529 | | 539 521 |
| pos | 12 | < > | 48 905 |
| case | 6 | | |
| proc | 4 | | |
| afun | 88 | | |
| parent | 332 | | |
| eparent | 13 189 | | |
| prep | 105 | | |
| p_lemma | 441 023 | | |
| p_tag | 2 604 | | |
| p_pos | 13 | | |
| p_case | 9 | | |
| p_afun | 88 | | |
| ep_lemma | 728 725 | | |
| ep_tag | 43 711 | | |
| ep_pos | 787 | | |
| ep_case | 1 282 | | |
| ep_afun | 1 654 | | |

Peážedka: Etika v tomto přehledu udává počet různých atributů / struktur (s, typů) v korpusu.

Citací informace:

Korpus jako zdroj dat:

Křen, M. – Oršák, V. – Čapka, T. – Čermáková, A. – Hrdáková, M. – Chlumská, L. – Jelišek, T. – Kovářková, D. – Petkevič, V. – Procházka, P. – Škoumalová, H. – Škrabal, M. – Truneček, P. – Vondříčka, P. – Zesina, A. SYN2015 reprezentativní korpus psané češtiny. Ústav českého národního korpusu FF UK, Praha 2015. Corponary z WWW <http://www.korpus.cz>.

Literatura:

Oršák, V. – Čermáková, A. – Křen, M. (2016). Nový koncept synchronních korpusů psané češtiny. Slovo a slovesnost, 77 (2), 89–101. ISSN 0037-7031.

Křen, M. – Oršák, V. – Čapka, T. – Čermáková, A. – Hrdáková, M. – Chlumská, L. – Jelišek, T. – Kovářková, D. – Petkevič, V. – Procházka, P. – Škoumalová, H. – Škrabal, M. – Truneček, P. – Vondříčka, P. – Zesina, A. (2014). SYN2015: Representative Corpus of Contemporary Written Czech. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'14), 2322–2328. Portorož: ELRA. ISBN 978-2-9311438-9-1. http://www.lrec-conf.org/proceedings/lrec2014/pdf/186_Paper.pdf

Obecné odkazy:

<http://www.korpus.cz>

Stručné info. vč. citování

Proč je třeba citovat korpusy

Jak číst informace o velikosti korpusu:

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

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

pozice

tokenizace

lemmatizace

desambiguace

Vyhledávání:

Typ dotazu: Základní

Dotaz:

klávesnice předchozí dotazy

Slovní tvar/slovo/word

Fráze

Lemma

CQL

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

konkordance, KWIC

Korpus: [syn2015](#) | Dotaz: [kočka](#) (8 507 výskytů) ▶ Promíchat: ✓

Výskytů: 8 507 | i.p.m.: 70,45 (vztaženo k celému korpusu) | ARF: 2 161,98 | Výsledek je promíchán 1 / 213 ▶▶▶

Výběr řádků: základní ▾

| | | | | |
|--------------------------|--|---|----------------|--------------------------------|
| <input type="checkbox"/> | | přečíst celičké vydání tak jako | kočka | vylizuje talířek s mlékem , |
| <input type="checkbox"/> | | Následující opatření si prosím na | kočce | předem nezkoušejte : dýchání z |
| <input type="checkbox"/> | | . Odvážnější kousky Pokud vaše | kočka | ve cvičení s klikrem najde |
| <input type="checkbox"/> | | uchu jsou varovnými signály – | kočka | si mohla do uší zanést |
| <input type="checkbox"/> | | na venkově žije se svými | kočkami | . Máš před sebou celý |
| <input type="checkbox"/> | | pěkně srandovní jméno pro lítající | kočku | . " „ Není to |
| <input type="checkbox"/> | | vhnrula do kuchyně . Jen | kočky | zmerčily mladou husičku , oči |
| <input type="checkbox"/> | | Týká se to psů , | koček | , ptáků i různých hlodavců |
| <input type="checkbox"/> | | ODRÁŽKA Máte v úmyslu pouštět | kočku | volně ven ? S prvním |
| <input type="checkbox"/> | | byť tak , aby se | kočce | líbil . Ale i pokud |
| <input type="checkbox"/> | | , totiž přibývá . Perská | kočka | Vhodné prostředí : byt , |
| <input type="checkbox"/> | | bengálku zamiluje . Britské krátkosrsté | kočky | Vzhled Zaoblené tvary těla a |
| <input type="checkbox"/> | | máte v bytě asi cizí | kočku | . Já si tohle rozhodně |

Zobrazení

KWIC/Věta

Korpus: [syn2015](#) | Dotaz: [kočka](#) (8 507 výskytů) ▶ Promíchat: ✓

Výskytů: 8 507 | i.p.m.: 70,45 (vztaženo k celému korpusu) | ARF: 2 161,98 | Výsledek je promíchán 1 / 213 ▶▶▶

Výběr řádků: základní ▾

| | | |
|--------------------------|--|--|
| <input type="checkbox"/> | | Sednout si na výsluní a přečíst celičké vydání tak jako kočka vylizuje talířek s mlékem , do posledního písmenka . |
| <input type="checkbox"/> | | Následující opatření si prosím na kočce předem nezkoušejte : dýchání z úst do úst , stimulaci dechu , srdeční masáž . |
| <input type="checkbox"/> | | Pokud vaše kočka ve cvičení s klikrem najde zalíbení , můžete přejít ke složitějším kouskům . |
| <input type="checkbox"/> | | Také časté potřásání hlavou a škrabání se v uchu jsou varovnými signály – kočka si mohla do uší zanést parazity nebo trpí ušní infekcí . |
| <input type="checkbox"/> | | Nevím , jestli ti v tomhle ohledu dokážu porozumět , protože koneckonců jsem jen stará ženská , která na venkově žije se svými kočkami . |
| <input type="checkbox"/> | | „ Jo ještě k tomu , Malcolm je pěkně srandovní jméno pro lítající kočku . ” |
| <input type="checkbox"/> | | Jen kočky zmerčily mladou husičku , oči se jim rozsvítily . |
| <input type="checkbox"/> | | Týká se to psů , koček , ptáků i různých hlodavců . |
| <input type="checkbox"/> | | ODRÁŽKA Máte v úmyslu pouštět kočku volně ven ? |
| <input type="checkbox"/> | | Majitelé koček se šikovnými rukama a kutilskými sklony mohou po návštěvě hobbymarketu sami kreativně zařídit byt tak , aby se kočce líbil . |
| <input type="checkbox"/> | | Perská kočka |

Korpusová nastavení

(Lemma, POS – **p**art **o**f **s**peech)

Korpus: syn2015 | Dotaz: kočka (8 507 výskytů) ▶ Promíchat: ✓

Výskytů: 8 507 | i.p.m.: 70,45 (vztaženo k celému korpusu) | ARF: 2 161,98 | Výsledek je promíchán 1 / 213 ▶▶▶

Výběr řádků: základní ▾

| | | | | |
|--------------------------|--|------------------------------------|-------------------|--------------------------------|
| <input type="checkbox"/> | | přečíst celičké vydání tak jako | kočka / kočka/N | vylizuje talířek s mlékem , |
| <input type="checkbox"/> | | Následující opatření si prosím na | kočce / kočka/N | předem nezkoušejte : dýchání z |
| <input type="checkbox"/> | | . Odvážnější kousky Pokud vaše | kočka / kočka/N | ve cvičení s klikrem najde |
| <input type="checkbox"/> | | uchu jsou varovnými signály – | kočka / kočka/N | si mohla do uší zanést |
| <input type="checkbox"/> | | na venkově žije se svými | kočkami / kočka/N | . Máš před sebou celý |
| <input type="checkbox"/> | | pěkně srandovní jméno pro lítající | kočku / kočka/N | . „ Není to |
| <input type="checkbox"/> | | vhrnula do kuchyně . Jen | kočky / kočka/N | zmerčily mladou husičku , oči |
| <input type="checkbox"/> | | Týká se to psů , | koček / kočka/N | , ptáků i různých hlodavců |
| <input type="checkbox"/> | | ODRÁŽKA Máte v úmyslu pouštět | kočku / kočka/N | volně ven ? S prvním |
| <input type="checkbox"/> | | bvt tak . abv se | kočce / kočka/N | líbil . Ale i pokud |

(Lemma, tag)

Korpus: syn2015 | Dotaz: kočka (8 507 výskytů) ▶ Promíchat: ✓ Nastavení bylo uloženo X

Výskytů: 8 507 | i.p.m.: 70,45 (vztaženo k celému korpusu) | ARF: 2 161,98 | Výsledek je promíchán 1 / 213 ▶▶▶

Výběr řádků: základní ▾

| | | | | |
|--------------------------|--|------------------------------------|----------------------------------|--------------------------------|
| <input type="checkbox"/> | | přečíst celičké vydání tak jako | kočka / kočka/NNFS1-----A----- | vylizuje talířek s mlékem , |
| <input type="checkbox"/> | | Následující opatření si prosím na | kočce / kočka/NNFS6-----A----- | předem nezkoušejte : dýchání z |
| <input type="checkbox"/> | | . Odvážnější kousky Pokud vaše | kočka / kočka/NNFS1-----A----- | ve cvičení s klikrem najde |
| <input type="checkbox"/> | | uchu jsou varovnými signály – | kočka / kočka/NNFS1-----A----- | si mohla do uší zanést |
| <input type="checkbox"/> | | na venkově žije se svými | kočkami / kočka/NNFP7-----A----- | . Máš před sebou celý |
| <input type="checkbox"/> | | pěkně srandovní jméno pro lítající | kočku / kočka/NNFS4-----A----- | . „ Není to |

Korpusová nastavení pro syn2015 X

Poziční atributy Struktury **Metainformace**

| | | | | | |
|---|--|--|--|---|---|
| <input type="checkbox"/> <#> <input type="checkbox"/> Počet tokenů <input type="checkbox"/> Vybrat vše | <input checked="" type="checkbox"/> <doc> <input type="checkbox"/> Pořadí dokumentu <input checked="" type="checkbox"/> doc.title <input type="checkbox"/> doc.subtitle <input checked="" type="checkbox"/> doc.author <input type="checkbox"/> doc.issue <input type="checkbox"/> doc.publisher <input type="checkbox"/> doc.pubplace <input type="checkbox"/> doc.pubyear <input type="checkbox"/> doc.first_published <input type="checkbox"/> doc.translator <input type="checkbox"/> doc.srclang <input type="checkbox"/> doc.authsex <input type="checkbox"/> doc.transsex <input type="checkbox"/> doc.txtype_group <input type="checkbox"/> doc.txtype | <input type="checkbox"/> <text> <input type="checkbox"/> text.section <input type="checkbox"/> text.section_orig <input type="checkbox"/> text.author <input type="checkbox"/> text.id <input type="checkbox"/> Vybrat vše | <input type="checkbox"/> <p> <input type="checkbox"/> p.type <input type="checkbox"/> p.id <input type="checkbox"/> Vybrat vše | <input type="checkbox"/> <s> <input type="checkbox"/> s.id <input type="checkbox"/> Vybrat vše | <input type="checkbox"/> <hi> <input type="checkbox"/> hi.rend <input type="checkbox"/> Vybrat vše |
|---|--|--|--|---|---|

Korpus: syn2015 | Dotaz: kočka (8 507 výskytů) ▶ Promíchat: ✓

Výskytů: 8 507 | i.p.m.: 70,45 (vztaženo k celému korpusu) | ARF: 2 161,98 | Výsledek je promíchán 1 / 213 ▶▶

Výběr řádků: základní ▾

| | | | |
|--------------------------|---|------------------------------------|---------------------------------|
| <input type="checkbox"/> | Konec světa & Hard-boiled Wonderland ↗ Murakami, Haruki | prečíst celíčké vydání tak jako | kočka /kočka/NNFS1-----A----- |
| <input type="checkbox"/> | Kočky ↗ Jones-Baade, Renate | Následující opatření si prosím na | kočce /kočka/NNFS6-----A----- |
| <input type="checkbox"/> | Kočky ↗ Jones-Baade, Renate | . Odvážnější kusky Pokud vaše | kočka /kočka/NNFS1-----A----- |
| <input type="checkbox"/> | Kočky ↗ Jones-Baade, Renate | uchu jsou varovnými signály – | kočka /kočka/NNFS1-----A----- |
| <input type="checkbox"/> | Páteční klub pletení ↗ Jacobs, Kate | na venkově žije se svými | kočkami /kočka/NNFP7-----A----- |
| <input type="checkbox"/> | Panoptikon ↗ Fagan, Jenni | pěkně srandovní jméno pro lítající | kočku /kočka/NNFS4-----A----- |
| <input type="checkbox"/> | Miláčci, máte se nač těšit! ↗ Welshman, Malcolm D. | vhmula do kuchyně . Jen | kočky /kočka/NNFS2-----A----- |
| <input type="checkbox"/> | Moje rodina ↗ X | Týká se to psů , | koček /kočka/NNFP2-----A----- |
| <input type="checkbox"/> | Kočky ↗ Jones-Baade, Renate | ODRÁŽKA Máte v úmyslu pouštět | kočku /kočka/NNFS4-----A----- |

Kompletní info o zdrojovém textu:

| | | | |
|------------------|--|---------------------|----------------------------|
| doc.title | Konec světa & Hard-boiled Wonderland | doc.subtitle | |
| doc.author | Murakami, Haruki | doc.issue | |
| doc.publisher | Odeon | doc.pubplace | Praha |
| doc.pubyear | 2008 | doc.first_published | 2008 |
| doc.translator | Jurkovič, Tomáš | doc.srclang | ja: japonština |
| doc.authsex | M: muž | doc.transsex | M: muž |
| doc.txtype_group | FIC: beletrie | doc.txtype | NOV: próza |
| doc.genre_group | X: neuvedeno | doc.genre | X: neuvedeno |
| doc.medium | B: kniha | doc.periodicity | NP: neperiodická publikace |
| doc.audience | GEN: obecné publikum | doc.isbnissn | 978-80-207-1287-5 |
| doc.biblio | Murakami, Haruki (2008): Konec světa & Hard-boiled Wonderland. Překlad: Jurkovič, Tomáš. Praha: Odeon. | doc.id | murak_konecsveta |
| text.section | | text.section_orig | |
| text.author | | text.id | murak_konecsveta:2 |
| p.type | normal | p.id | murak_konecsveta:2:1920 |

Úkol na příště:

Prostudovat www stránky ÚČNK

Umět odpovědět na otázky:

1. Co je to korpusu?
2. Co je to Český národní korpus?
3. Jaké typy korpusů máme k dispozici?
4. Co to znamená, když řeknu, že korpus má 100 milionů slov?
5. Jak komunikujeme s korpusem (jak jej můžeme využívat pro lingvistickou práci)?
6. Jak můžeme vyhledat v korpusu výskyt slova, jak se se zobrazí v korpusu výskyt slova a co můžeme se zobrazenými výskyty dále dělat?
7. Jak můžeme vyhledat v korpusu všechna slova, která mají společnou vlastnost, že jsou tvary jednoho základního tvaru?

8. Jak můžeme v korpusu vyhledat všechny tvary na rovině gramatické abstrakce (třeba podstatné jména rodu ženského ve 3. pádě, nebo slovesa v přítomném čase v první osobě)?

A připravit si otázky, na něž byste rádi znali odpověď (souvisí s korpusy!!)