

Úvod do korpusové lingvistiky 5



PRAXE V ČESKÉM PROSTŘEDÍ

Elektronicky přístupný skrze **korpusové manažery**



- Klient – server
- Webové rozhraní

BONITO



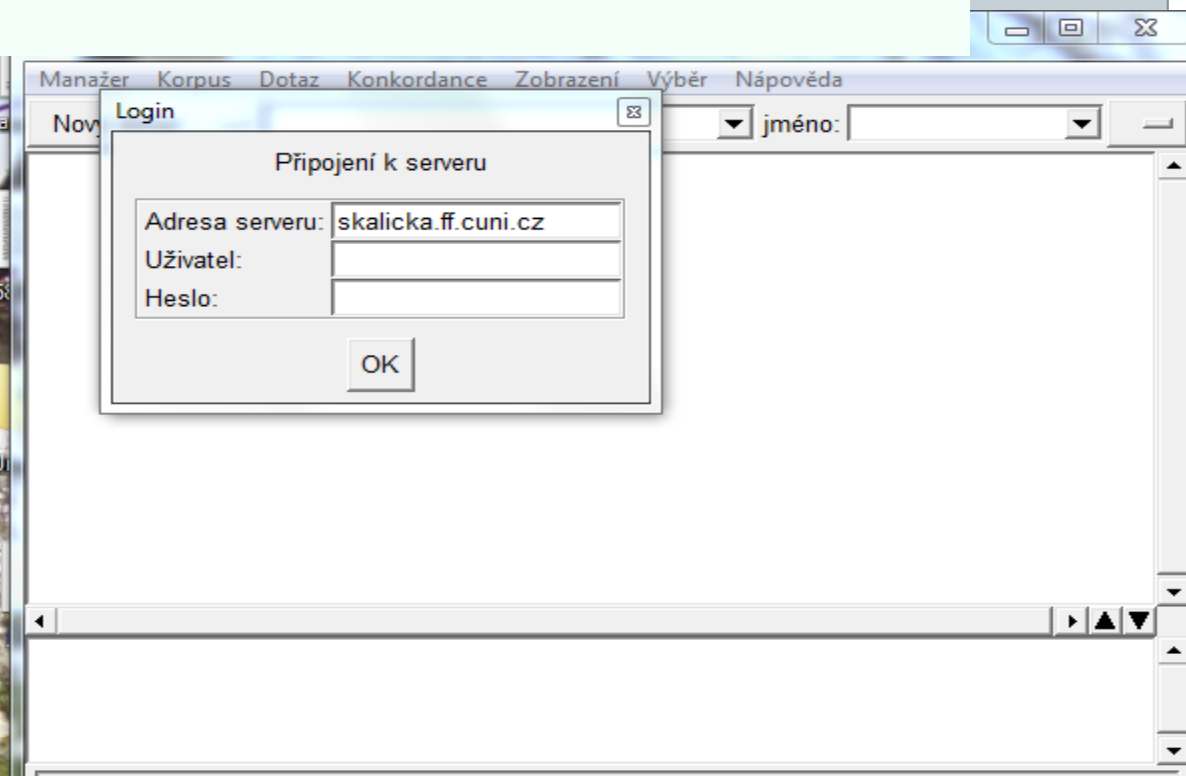
- **Bonito**

Program Bonito vznikl v Centru zpracování přirozeného jazyka FI MU a jeho autorem je Pavel Rychlý.

Zde je návod na instalaci programu *Bonito* v prostředí operačního systému **Windows** (95, 98, ME, 2000, XP, Vista, 7):

- Nejprve si na svůj počítač stáhněte následující soubor:

[bonito-install.exe](#) (1,5 MB)



NoSketch Engine

● Vyzkoušejte si nové webové rozhraní



- https://korpus.cz/corpora/run.cgi/first_form

The screenshot shows a web browser window with the URL https://korpus.cz/corpora/run.cgi/first_form. The page header includes the NoSketch Engine logo and navigation links: "Nejnavštěvovanější" and "Jak začít". Below the header, the user information is displayed: "Uživatel: osolsobe", "Korpus: [syn2010](#)", "Popis: [Synchronní reprezentativní korpus](#)", and "Velikost: 121 667 413 pozic".

The main content area is divided into a left sidebar and a central search form. The sidebar contains links for "Nový dotaz", "Seznam slov", "Pokročilá nastavení" (with sub-links for "Kontext" and "Subkorpus"), and "Uživatel" (with a link for "Změna hesla").

The central search form includes the following fields and buttons:

- Korpus:** A text input field containing "syn2010".
- Typ dotazu:** A dropdown menu set to "Základní".
- Dotaz:** An empty text input field with a keyboard icon to its right.
- Buttons:** "Hledat" and "Smazat formulář".

Doporučujeme



- Varianta klient-server (**bonito**) není v současné době již dále udržovaná
- Varianta **NoSketch Engine** prošla v poslední době úpravami a bude i nadále udržovaná
- Varianta klient-server má velmi dobrý uživatelský manuál
- Varianta No Sketch Engine má dosud k dispozici pouze manuál ke starší (neaktuální) verzi
- **Křen, Michal: *Co je nového v ČNK II. KORPUS – GRAMATIKA – AXIOLOGIE* 7/ 2013, 98-100.**

Základní termíny a funkce manažeru



- Příkazová okna
- KWIC (**K**ey **W**ord **I**n **C**ontext)
- Konkordance / konkordanční seznam
- Frekvenční seznam
- Filtr
- Zobrazení
- Uložení

Volba korpusu



Nejnavštěvovanější Jak začít

NoSketch Engine

Uživatel: osolsobe Korpus: syn2010 Popis: Synchronní reprezentativní korpus Velikost: 121 667 413 pozic ²

[Nový dotaz](#)

[Seznam slov](#)



Pokročilá nastavení:

[Kontext](#)

[Subkorpus](#)

Uživatel:

[Změna hesla](#)

Korpus:

syn2010

Typ dotazu:

Dotaz:

Hledat

Smazat formulář

▼ Synchronní psané korpusy

▼ řada SYN

[syn](#)

[syn2010](#)

[syn2009pub](#)

[syn2006pub](#)

[syn2005](#)

[syn2000](#)

▼ specializované

[czesl-plain](#)

[ksk-dopisy](#)

[link](#)

[orw-mte](#)

[orwell](#)

[skript2012](#)

▼ ke slovníkům

[fsc2000](#)

▼ Synchronní mluvené korpusy

▶ řada ORAL

▶ specializované

▼ Diachronní korpusy

[diakorp](#)

Volba vyhledávání



- Základní / Lemma / Fráze / Slovní tvar / Podřetězec /CQL

NoSketch Engine

Uživatel: osolsobe Korpus: [syn2010](#) Popis: [Synchronní reprezentativní korpus](#) Velikost: 121 667 413 pozic [?]


[Nový dotaz](#)
[Seznam slov](#) [?]

Pokročilá nastavení:
[Kontext](#) [?]
[Subkorpus](#) [?]

Uživatel:
[Změna hesla](#)

Korpus:

Typ dotazu: [?]

Dotaz: 

Slovní tvar (word)



- Textová slova (**češu**)

Nejnavštěvovanější Jak začít

NoSketch Engine

Uživatel: osolsobe Korpus: syn2010 Popis: Synchronní reprezentativní korpus Velikost: 121 667 413 pozic [?]

[Nový dotaz](#)
[Seznam slov](#) [?]

Pokročilá nastavení:
[Kontext](#) [?]
[Subkorpus](#) [?]

Uživatel:
[Změna hesla](#)

Korpus:

Typ dotazu: ▼

Slovní tvar:

Konkordanční seznam ve formě **KWIC** (Key Word In Context)



- Vyhledá všechny kontexty výskytu hledaného slovního tvaru

NoSketch Engine

Uživatel: osolsobe Korpus: syn2010 Popis: Synchronní reprezentativní korpus Velikost: 121 667 413 pozic ² Výskytů: 7

Nový dotaz
Seznam slov
Uložit
Zobrazení
 Vlastní
 KWIC
 Věta
Třídění
 Vlastní

Výskytů: 7 i.p.m.: 0,06 (vztaženo k celému korpusu) | ARF: 4

Jsem statečná žena	víš jak moc rád tě češu /česat/VB-S---1P-AA--- . Kdyby sis mě vzala	
Pavučina	" Můžeš být klidný , češu /česat/VB-S---1P-AA--- se svým vlastním hřebenem !	
Kniha radosti	když jsem vytahovala vysavač . Češu /česat/VB-S---1P-AA--- se naposledy , říkala jsem	
Lidé a země, č. 6/2008	. Obouvám si boty , češu /česat/VB-S---1P-AA--- se . . . každá	
Lidové noviny, 4. 12. 2006	. Zatímco běžně perskou kočku češu /česat/VB-S---1P-AA--- denně tak deset minut ,	
Deníky Bohemia, 12. 4. 2007	a půl hodiny se třeba češu /česat/VB-S---1P-AA--- , " překvapila tanečnice ,	
STYL, č. 41/2008	jde . Sama se i češu /česat/VB-S---1P-AA--- a líčím . . .	

Lemma – základní tvar – systémové slovo



- Lemma (**česat**)

Nejnavštěvovanější Jak začít

NoSketch Engine

Uživatel: osolsobe Korpus: syn2010 Popis: Synchronní reprezentativní korpus Velikost: 121 667 413 pozic [?]

[Nový dotaz](#)
[Seznam slov](#) [?]

Pokročilá nastavení:
[Kontext](#) [?]
[Subkorpus](#) [?]

Uživatel:
[Změna hesla](#)

Korpus:

Typ dotazu:

Lemma:

Konkordanční seznam ve formě KWIC (Key Word In Context)



- Výskytů tvarů (word) hledaného lemmatu je **291**

Nejnavštěvovanější Jak začít

NoSketch Engine

Hledat

Uživatel: osolsobe Korpus: [syn2010](#) Popis: [Synchronní reprezentativní korpus](#) Velikost: 121 667 413 pozic² Výskytů: 291

Nový dotaz
[Seznam slov](#)

Uložit
Zobrazení
[Vlastní](#)
[KWIC](#)
[Věta](#)

Třídění
[Vlastní](#)
[Vlevo](#) | [Vpravo](#)
[KWIC](#)
[Reference](#)
[Promíchat](#)

[Vzorek](#)
Filtr

Výskytů: 291 i.p.m.: 2,39 (vztaženo k celému korpusu) | ARF: 123

strana ze 15 [další](#) | [poslední](#)

Ruská sekce	Novgorodu prý omdlela , když	česala /česat/VpFS---3R-AA---	jablka , přímo na žebříku	
Fantom chrámu	rukávu hřeben a začal si	česat /česat/Vf-----A----	kníry a vousy . Věnoval	
Za trest	, " řekl otec a	česal /česat/VpMS---3R-AA---	si přítom husté , kudrnaté	
Dívčí hrob	, jako by mu je	česal /česat/VpMS---3R-AA---	samotný Dean Stillwell . Mělo	
Splátka růží	tam všechny dívky , které	česaly /česat/VpFP---3R-AA---	a mykaly čerstvě obarvenou vlnu	
Železná brána	volné kůže . Přestala si	česat /česat/Vf-----A----	vlasý a muselo se jí	
Učedník	, balzamuje , obléká ,	češe /česat/VB-S---3P-AA---	, líčí a tak podobně	
Učedník	hezkých pár dní nemyla ani	nečesala /česat/VpFS---3R-NA---	. Nedotýkala se Korsaka ,	
Chladnokrevně	panenkou , kterou myla a	česala /česat/VpNP---3R-AA---	a líbala , někdy od	
Konečné řešení	chlapci . Umývala ho a	česala /česat/VpNP---3R-AA---	. Krmila ho , šatila	
Muka dospívání aneb Těžký hypochondr v pubertě	zuby pravidelně . Denně se	češeš /česat/VB-S---2P-AA---	, denně si i čistí	
Pětka	už si sundal přilbu a	česal /česat/VpMS---3R-AA---	se před zpětným zrcátkem .	
Bílá ruka a poklad hradu Handštejna	výšky , a stále si	česal /česat/VpIS---3R-AA---	pečlivě pěšinku , jak byl	
Eragon	jméno broukala , když mu	česala /česat/VpFS---3R-AA---	vlasý . . . Stín	

Definování zobrazení hodnot lemmatu, tagu a díla pro KWIC



- Zobrazení

Zobrazení

Vlastní

KWIC

Věta

Možn. zobrazení

Atributy	Struktury	Reference
<input checked="" type="checkbox"/> word	<opus>	Token number
<input checked="" type="checkbox"/> lemma	<doc>	Document number
<input checked="" type="checkbox"/> tag	<s>	opus.nazev
<input type="checkbox"/> lc		opus.autor
<input type="checkbox"/> pos		opus.nakladatel
<input type="checkbox"/> k		opus.mistovyd
<input type="checkbox"/> g		opus.rokvyd
<input type="checkbox"/> c		opus.isbnissn
		opus.preklad
		opus.srlang
		opus.btype_group
		opus.btype
		opus.genre

Zobrazit atributy

u všech tokenů

pouze u KWIC

Možnosti zjištění frekvenční distribuce



- Frekvenční distribuce

- Frekv. distribuce
 - Vlastní
 - Značky
 - Slovní tvary
 - Dokumenty
 - Typy textu

Víceúrovňová frekvenční distribuce

Frekvenční limit:

<input checked="" type="radio"/> první úroveň	<input type="radio"/> druhá úroveň	<input type="radio"/> třetí úroveň
Atribut: <input type="text" value="word"/>	Atribut: <input type="text" value="word"/>	Atribut: <input type="text" value="word"/>
Nerozlišovat velikost <input type="checkbox"/>	Nerozlišovat velikost <input type="checkbox"/>	Nerozlišovat velikost <input type="checkbox"/>
<input type="text" value="4L"/> <input type="text" value="3L"/> <input type="text" value="2L"/> <input type="text" value="1L"/>	<input type="text" value="4L"/> <input type="text" value="3L"/> <input type="text" value="2L"/> <input type="text" value="1L"/>	<input type="text" value="4L"/> <input type="text" value="3L"/> <input type="text" value="2L"/> <input type="text" value="1L"/>
Pozice: <input type="text" value="KWIC"/>	Pozice: <input type="text" value="KWIC"/>	Pozice: <input type="text" value="KWIC"/>
(Node) začít od [?] :	(Node) začít od [?] :	(Node) začít od [?] :
<input type="text" value="slova KWIC nejvíce vlevo"/>	<input type="text" value="slova KWIC nejvíce vlevo"/>	<input type="text" value="slova KWIC nejvíce vlevo"/>

Frekvenční distribuce slovních tvarů



Frekvenční distribuce

Frekvenční limit:

Celkem: 40 (1 str.)

	<u>word</u>	<u>Frekvence</u>	
1.	p/n česat	72	
2.	p/n česala	42	
3.	p/n češe	38	
4.	p/n česal	33	
5.	p/n češou	15	
6.	p/n česaly	10	
7.	p/n češeme	8	
8.	p/n česali	7	
9.	p/n nečešeme	7	
10.	p/n češu	6	
11.	p/n nečešou	4	
12.	p/n nečesala	4	
13.	p/n nečesal	4	
14.	p/n Česat	3	
15.	p/n nečeše	3	
16.	p/n nečesat	3	
17.	p/n Nečesal	3	
18.	p/n češeš	2	
19.	p/n češ	2	
20.	p/n česány	2	
21.	p/n česalo	2	
22.	p/n Česal	2	
23.	p/n Nečesej	2	
24.	p/n češete	1	
25.	p/n česán	1	

češou



Nejnavštěvovanější Jak začít

NoSketch Engine

Hle

Uživatel: osolsobe Korpus: syn2010 Popis: Synchronní reprezentativní korpus Velikost: 121 667 413 pozic ² Výskytů: 15

[Nový dotaz](#)

[Seznam slov](#)

[Uložít](#)

[Zobrazení](#)

[Vlastní](#)

[KWIC](#)

[Věta](#)

[Třídění](#)

[Vlastní](#)

[Vlevo](#) | [Vpravo](#)

[KWIC](#)

[Reference](#)

[Promíchat](#)

[Vzorek](#)

Výskytů: 15 i.p.m.: 0,12 (vztaženo k celému korpusu) | ARF: 6

Zóna Berlín	listí . Ted' si navzájem češou /česat/VB-P---3P-AA--- vlasy . Vplétá mu do	
Vévodkyně a kuchařka	lidé z bud ho přesto češou /česat/VB-P---3P-AA--- a někde jinde za pár	
Křehké knoflíky: Předměty - Jídlo - Pokoje	. Světnice , kde se češou /česat/VB-P---3P-AA--- slepice a peří a zralý	
Stálo to za hovno..., ale aspoň byla sranda	nejvíc připomínal to , co češou /česat/VB-P---3P-AA--- některé matky děťátkům , když	
Zločin pozdvížení	, jak se umývají , češou /česat/VB-P---3P-AA--- , svlékají a oblékají ,	
Jablko se kouše	dvě sedí u balvanu a češou /česat/VB-P---3P-AA--- si vlasy ; vpředu stojí	
Lidé a země, č. 10/2008	možný . Sběrači tedy postupně češou /česat/VB-P---3P-AA--- načervenalé bobule , jimž se	
Poznej sám sebe i druhé	Lidé , kteří si takto češou /česat/VB-P---3P-AA--- vlasy , se většinou snaží	
Poznej sám sebe i druhé	Lidé , kteří si takto češou /česat/VB-P---3P-AA--- vlasy , mají korektní ,	
Typ a účes	přirozeným ženám , které si češou /česat/VB-P---3P-AA--- vlasy jen prsty , a	
Typ a účes	a vlasy se pak snáze češou /česat/VB-P---3P-AA--- . Roztoky pro foukanou jsou	
Typ a účes	vlasy sestřihané stejně dlouze a češou /česat/VB-P---3P-AA--- se od spirálky na temeni	
Typ a účes	, většinou se však špatně češou /česat/VB-P---3P-AA--- a upravují . Všechny kadeře	
Encyklopedie kachlů v Čechách, na Moravě a ve Slezsku	dva soukenické knapy , jak češou /česat/VB-P---3P-AA--- štětkami sukno po zvalchování .	
Magazín DNES, č. 9/2005	, se všelijak ličí , češou /česat/VB-P---3P-AA--- , nastavují . Co děláte	

Filtry



- Na pozici KWIC <0,0> pouze tvary prézentu

Filtr konkordanci

Filtr: pozitivní negativní

Vybraný token: první poslední

Rozsah hledání: od do včetně KWIC

Typ dotazu:

CQL:

Implicitní atribut: word [Popis morfologických značek](#)

Pouze tvary [tag="VB.*"]



Učedník	, balzamuje , obléká ,	češe /česat/VB-S---3P-AA---I	, líčí a tak podobně
Muka dospívání aneb Těžký hypochondr v pubertě	zuby pravidelně . Denně se	češeš /česat/VB-S---2P-AA---I	, denně si i čistí
Nová láska na obzoru	půjdu . Proč se pořád	češeš /česat/VB-S---2P-AA---I	? Vždyť si vyrveš všechny
Paní jezera	Ve vzduchu tančí hřeben a	češe /česat/VB-S---3P-AA---I	její havraní kadeře . Yennefer
Gymnázium	děje , pro kterého se	češe /česat/VB-S---3P-AA---I	zelené zlato , aby se
Zapomenuté vzpomínky	košili a před zrcadlem si	češe /česat/VB-S---3P-AA---I	vlasý . " Jsi krásná
Zelená míle	Brad Dolan . Neustále si	češe /česat/VB-S---3P-AA---I	vlasý , stejně jako Percy
Srdceryvné dílo ohromujícího génia - kniha	batikovaný kalhoty a zásadně se	nečešou /česat/VB-P---3P-NA---I	. Většinou je přes čtyřicet
Srdceryvné dílo ohromujícího génia - kniha	jak chodí , jak se	češe /česat/VB-S---3P-AA---I	, jaká dělá gesta -
O lásce a jiných běsech	sedí u toaletního stolku a	češe /česat/VB-S---3P-AA---I	se jen tak pro nikoho
Zóna Berlín	listí . Ted' si navzájem	češou /česat/VB-P---3P-AA---I	vlasý . Vplétá mu do
Ukolébavka	. Chytí další hrst a	češe /česat/VB-S---3P-AA---I	, tupíruje , drbe ,
Vévodkyně a kuchařka	sad , v němž se	češe /česat/VB-S---3P-AA---I	ovoce , aby se svázelo
Vévodkyně a kuchařka	lidé z bud ho přesto	češou /česat/VB-P---3P-AA---I	a někde jinde za pár
Chlapectví	že si pomáduje vlasý a	češe /česat/VB-S---3P-AA---I	se na patku . On
Kainova kniha	plešatí a blond'até vlasý si	češe /česat/VB-S---3P-AA---I	dopředu a pomáduje si je
Jsem statečná žena	víš jak moc rád tě	češu /česat/VB-S---1P-AA---I	. Kdyby sis mě vzala
Osobní korespondence	frajer ale je hrozně pěkněj	češe /česat/VB-S---3P-AA---I	se na patku je to
Utrpení oddaného Všiváka	se oblékáme , stejně se	češeme /česat/VB-P---1P-AA---I	- - ve stejný den
Andělé pustiny	pouští v koupelně vodu a	češe /česat/VB-S---3P-AA---I	se všemi těmi hřebeny ,

Uložení



- Možnost uložení a práce of-line

Uložit konkordanci ?

Uložit konkordanci jako: Text XML CSV

Připojit hlavičku:

Include line numbers:

Zarovnat KWIC:

Uložit řádky: od do

Uložit konkordanci

Uložení do textového formátu



- Hlavička obsahuje informace o korpusu, s nímž pracujeme a o dotazu, přes nějž jsme získali uložená data

```
# Corpus: syn2010
# Hits: 89
# Relative frequency: 0.73 ((vztaženo k celému korpusu))
# ARF: 39.96
# Query: word,[lemma="česat"] 291
# Positive filter: 0 0 1 [tag="VB.*"] 89

Učedník , balzamuje , obléká , < češe /česat/VB-S--3P-AA--I > , ličí a tak podobně

Muka dospívání aneb Těžký hypochondr v pubertě zuby pravidelně . Denně se < češeš /česat/VB-S---
2P-AA--I > , denně si i čistí
Nová láska na obzoru půjdu . Proč se pořád < češeš /česat/VB-S--2P-AA--I > ? Vždyť
si vyrveš všechny
Paní jezera Ve vzduchu tančí hřeben a < češe /česat/VB-S--3P-AA--I > její havrani
kadeře . Yennefer
Gymnázium děje , pro kterého se < češe /česat/VB-S--3P-AA--I > zelené zlato , aby se

Zapomenuté vzpomínky košili a před zrcadlem si < češe /česat/VB-S--3P-AA--I > vlasy . "
Jsi krásná
Zelená míle Brad Dolan . Neustále si < češe /česat/VB-S--3P-AA--I > vlasy , stejně jako
Percy
Srdceryvné dílo ohromujícího génia - kniha batikovaný kalhoty a zásadně se < nečešou /česat/VB-P--3P-
NA--I > . Většinou je přes čtyřicet
```

Manuály k variantám korpusového manažeru



- Klient-server:

<http://ucnk.ff.cuni.cz/bonito/index.php>

- Webové rozhraní:

http://ucnk.ff.cuni.cz/doc/Bonito2_manual.pdf

Manuál korpusového manažeru Bonito

Marie Kopřivová
Jan Kocek

The screenshot shows the Bonito corpus manager interface. The window title is "Bonito". The menu bar includes "Manažer", "Korpus", "Dotaz", "Konkordance", "Zobrazení", "Výběr", and "Nápověda". The main area displays a list of concordance results for the query "[lemma='korpus']" in the corpus "syn2000". The results are organized into columns: opus, text, lemma, and context. The word "korpus" is highlighted in red in the lemma column. The interface includes a search bar, a corpus selection dropdown, and a status bar at the bottom showing "Zobrazeno: 1+100/276 (36%) Řádek: 7 Vybráno: 1".

Annotations on the right side of the image point to various elements:

- dotazový řádek
- výběr korpusu
- pojmenování dotazu
- konkordanční řádek
- označený konkordanční řádek
- vyhledaný výraz - KWIC (key word in context)
- konkordanční seznam
- kód jednoznačně identifikující text
- rozšíření kontextu vyhledaného výrazu
- otazový řádek

Začínáme s Bonitem 2 //Word Sketch Engine//

Východisko

Bonito 2 je internetový program, který lze použít na zpracování korpusu libovolného jazyka, je-li tento korpus označován vhodným způsobem.

Bonito 2 má řadu funkcí, z nichž základní jsou:

konkordancer (velmi rychlý a vysoce funkční)
program Word Sketch (viz dále).

Více informací o programu Word Sketch naleznete na [Kilgarriff et al 2004 in Proc EURALEX](#).

Od ledna 2014 KonText



Nový portál ČNK

Vážení uživatelé ČNK,
dne 28. 1. 2014 byl spuštěn portál pro práci s korpusovými daty. Je umístěn na adrese www.korpus.cz a jeho smyslem je inkorporovat všechny nástroje a informace pro práci s našimi korpusy. Hlavní novinkou je zcela nové rozhraní pro práci s korpusy, které jsme pojmenovali **KonText**.

Tým projektu ČNK

https://kontext.korpus.cz/run.cgi/first_form



KonText Park SyD Morfio KWords Wiki

kon text

Dotaz Subkorpusy | Uložit Konkordance Filtr Frekv. distribuce Kolokace Možn. zobrazení Nápověda

Hledat v korpusu

Korpus:

syn

Typ dotazu:

CQL

CQL:

[vložit tag](#) [vložit "within"](#) [klávesnice](#)

Implicitní atribut: word [Popis morfologických značek](#)

Specifikovat kontext ↓

Specifikovat dotaz podle metainformací ↓

Hledat

Vymazat formulář

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



- **Manuál práce s korpusovým rozhraním**
- **Funkce rozhraní KonText**
- **Přehled základních pojmů korpusové lingvistiky**
- **Jaké korpusy zpřístupňuje Český národní korpus?**
- **Seznamy zdrojů a zkratek**