Úvod do korpusové lingvistiky 5

PRAXE V ČESKÉM PROSTŘEDÍ

Elektronicky přístupný
skrze korpusové manažery

- Klient – server
- Webové rozhraní

BONITO

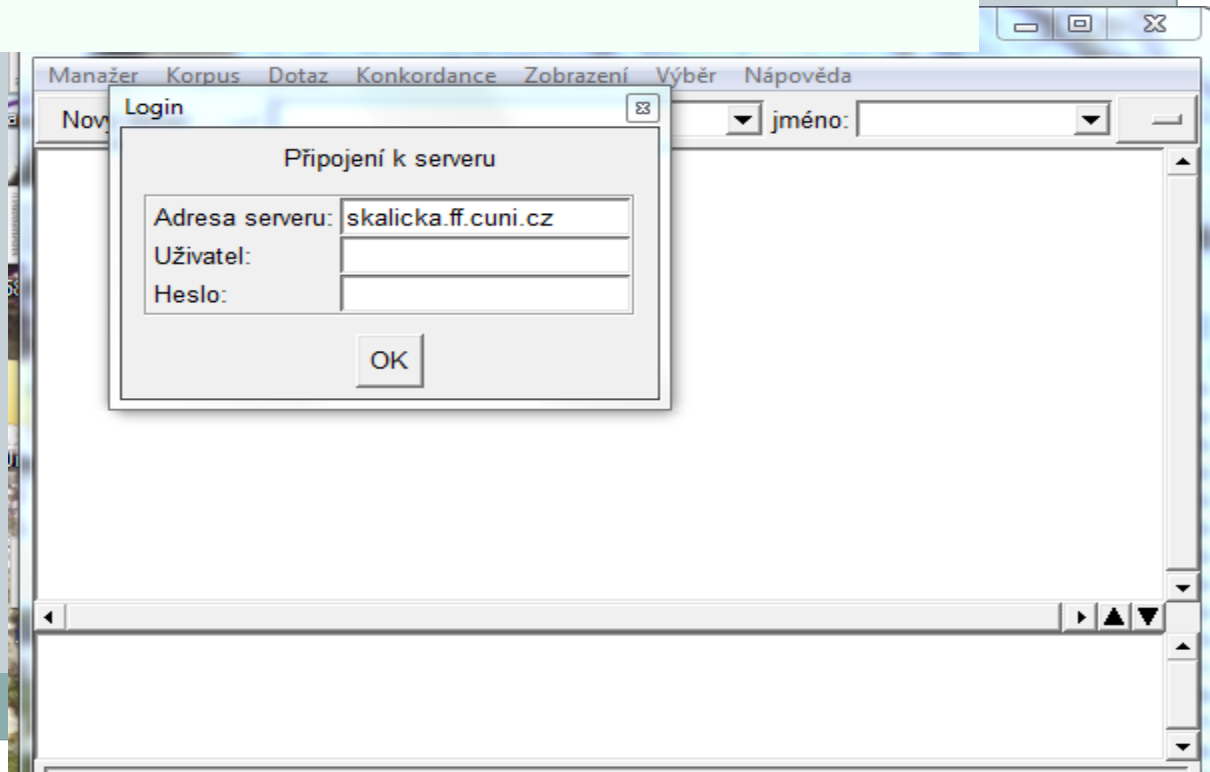
- Bonito

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:

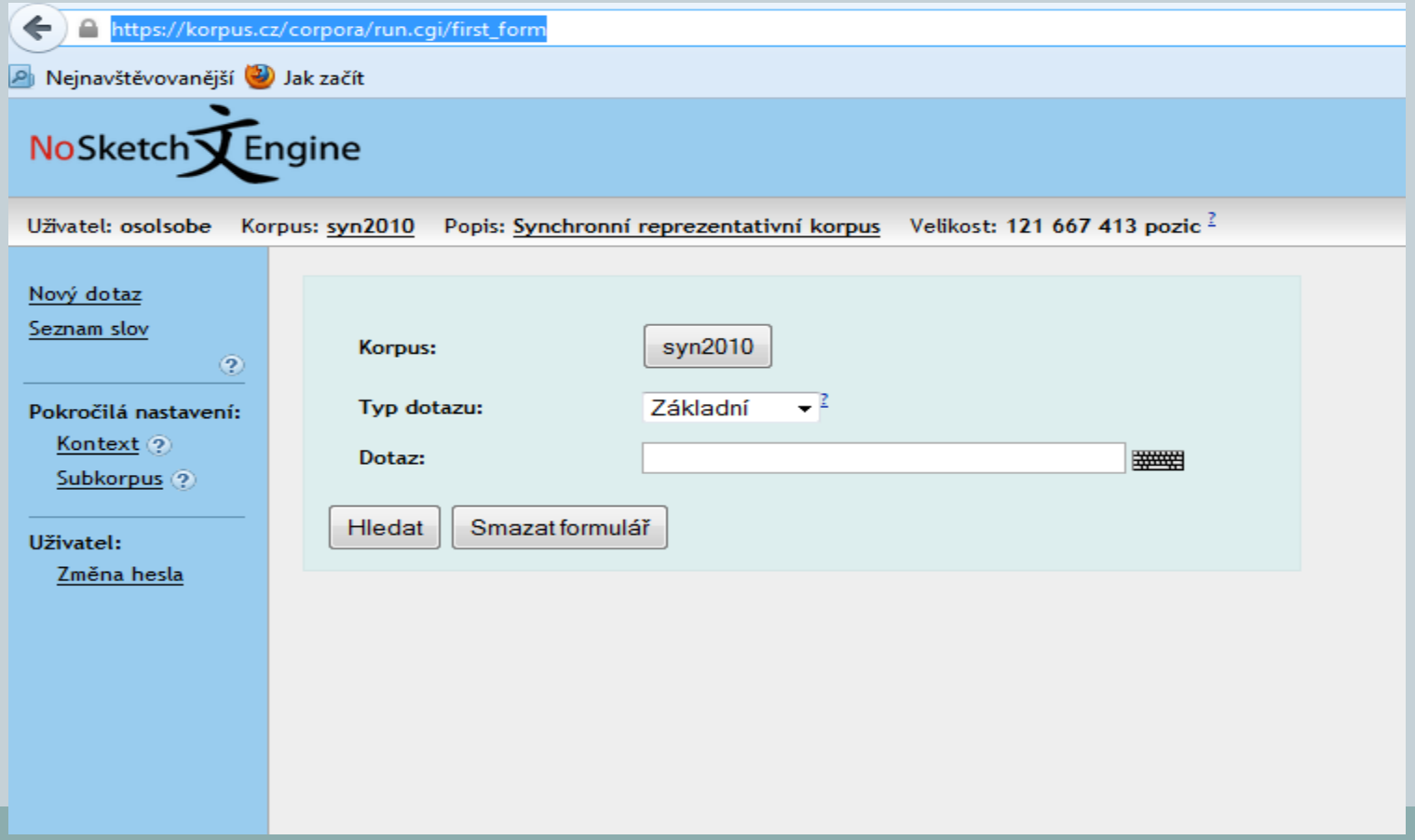
  bonito-install.exe (1,5 MB)



NoSketch Engine

- Vyzkoušejte si nové webové rozhraní

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



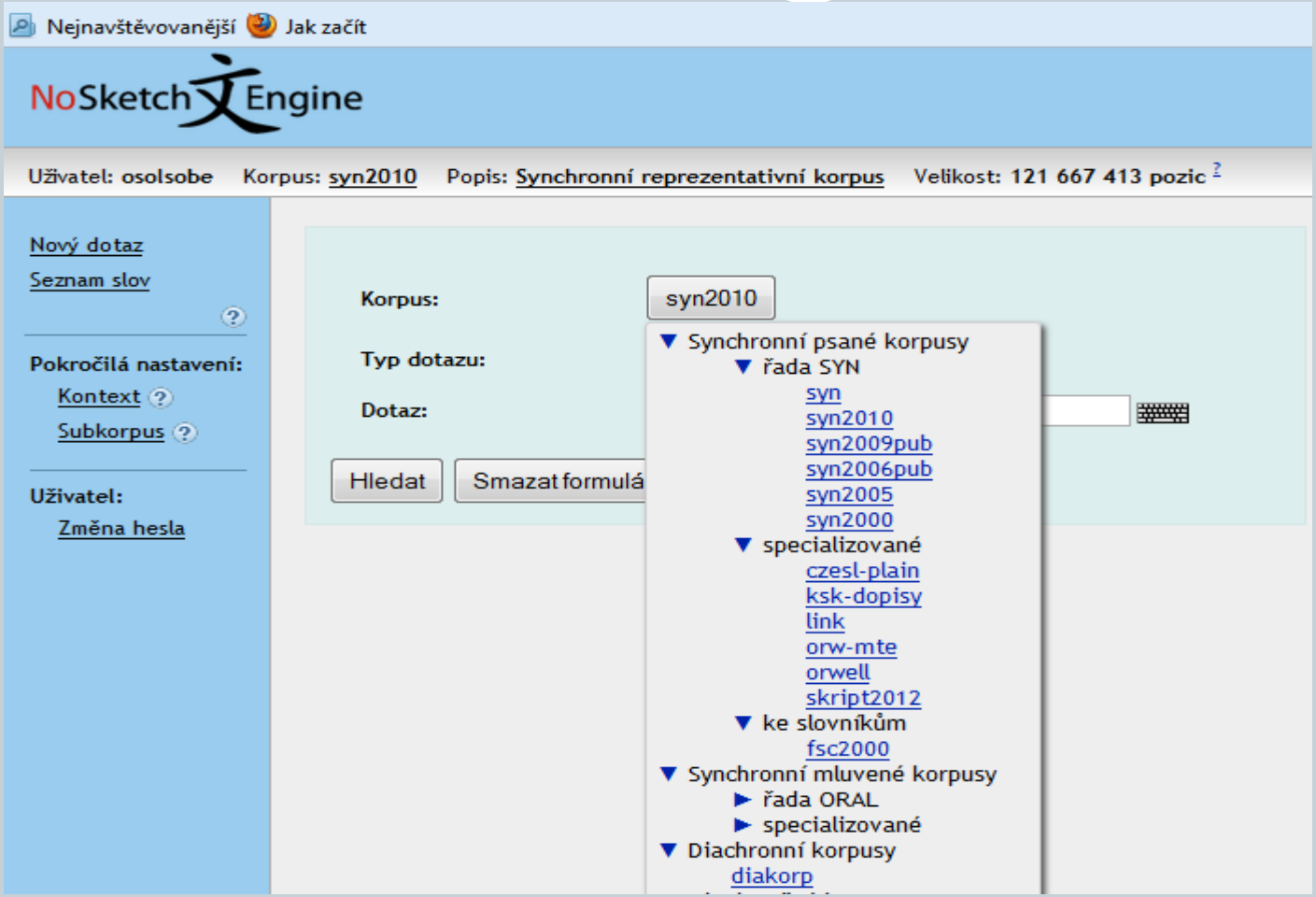
Doporučujeme

- 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

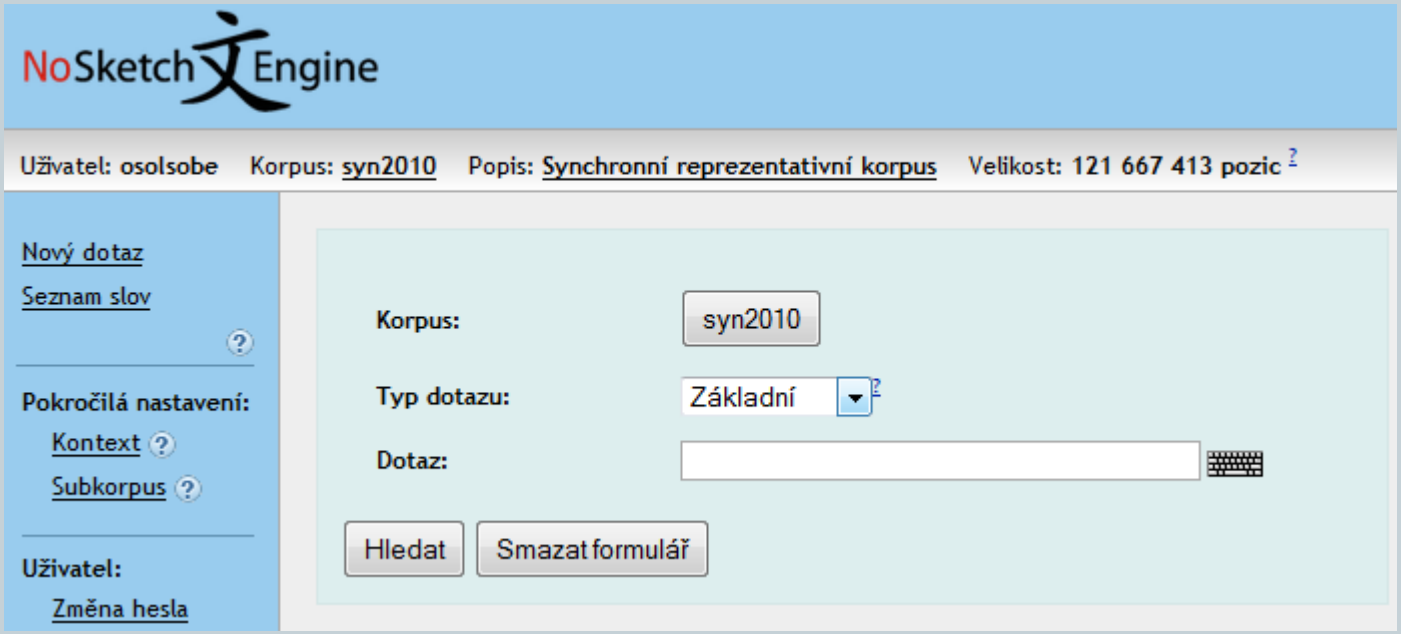
- Příkazová okna
- KWIC (Key Word In Context)
- Konkordance / konkordanční seznam
- Frekvenční seznam
- Filtr
- Zobrazení
- Uložení

Volba korpusu



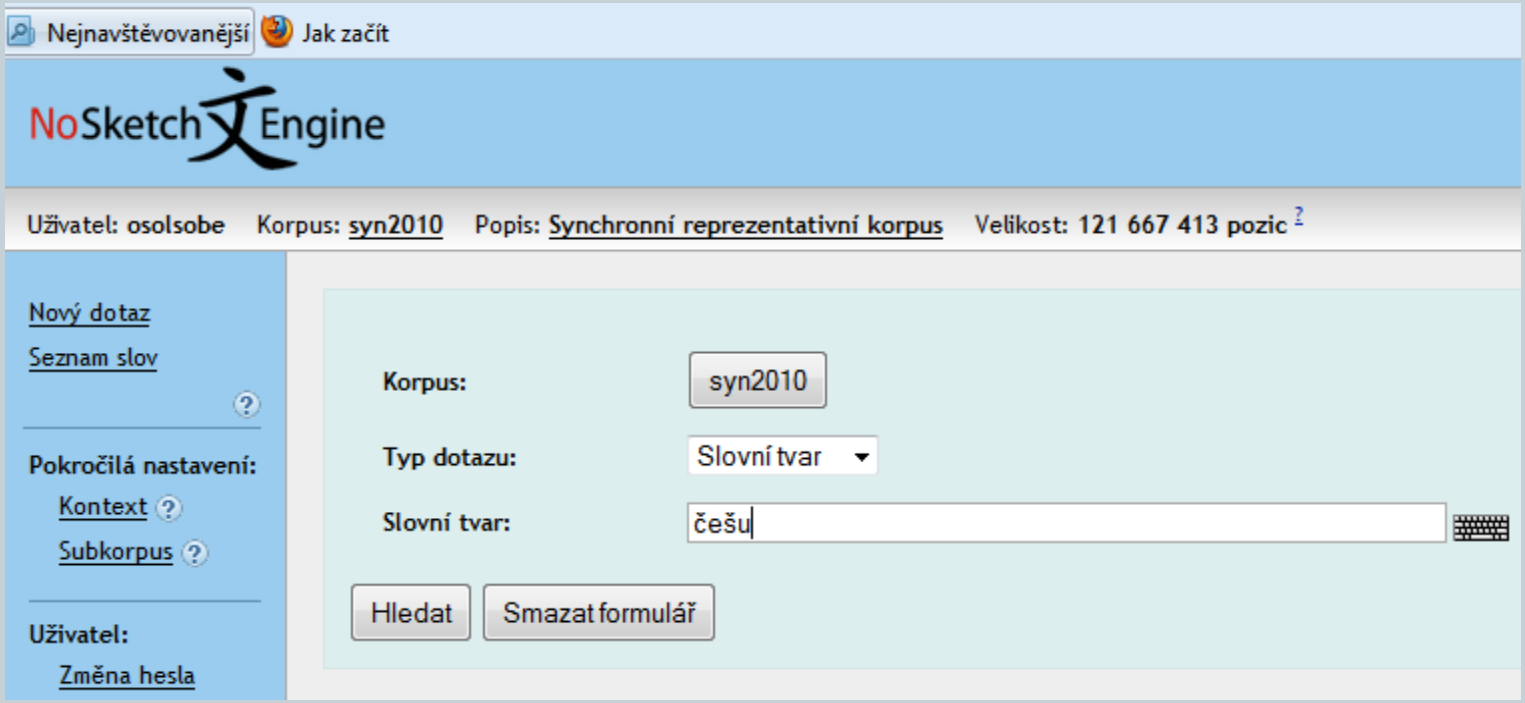
Volba vyhledávání

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



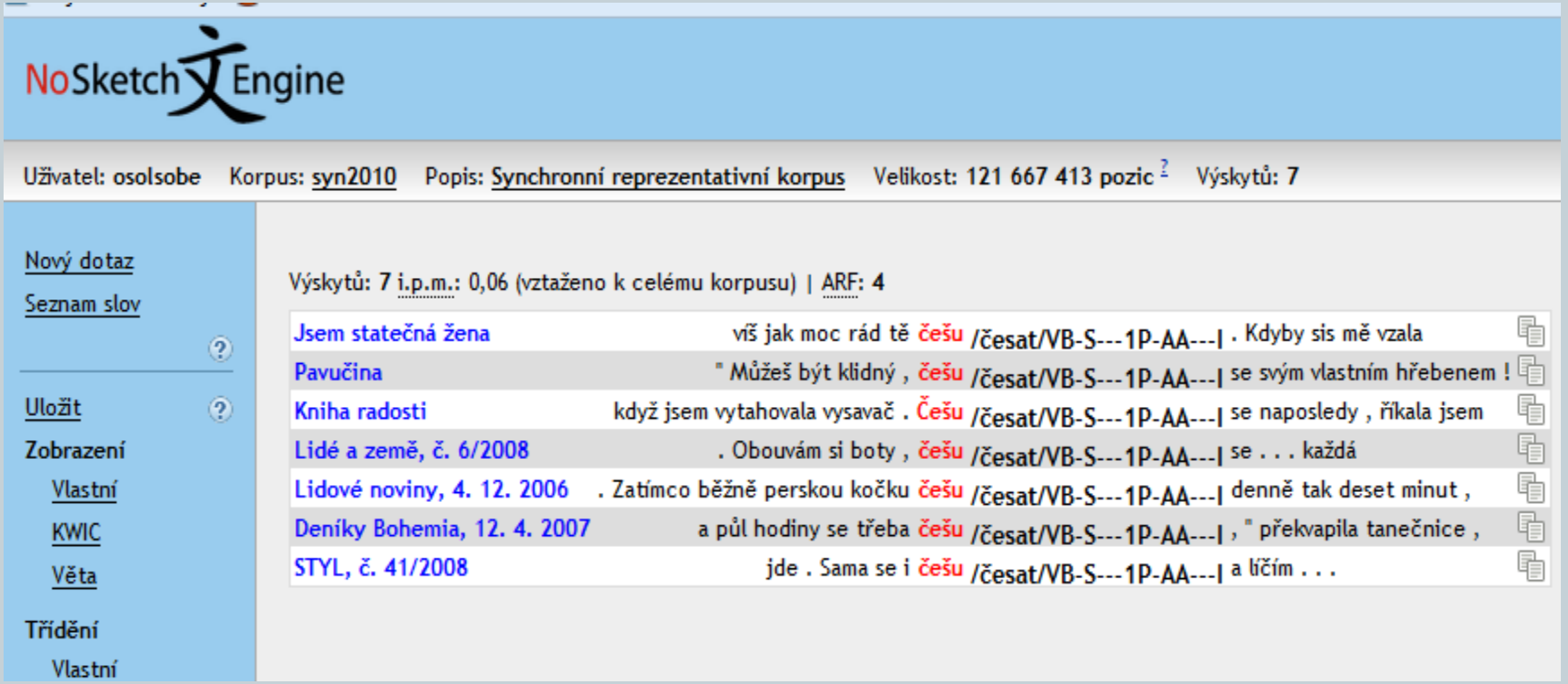
Slovní tvar (word)

- Textová slova (***češu***)



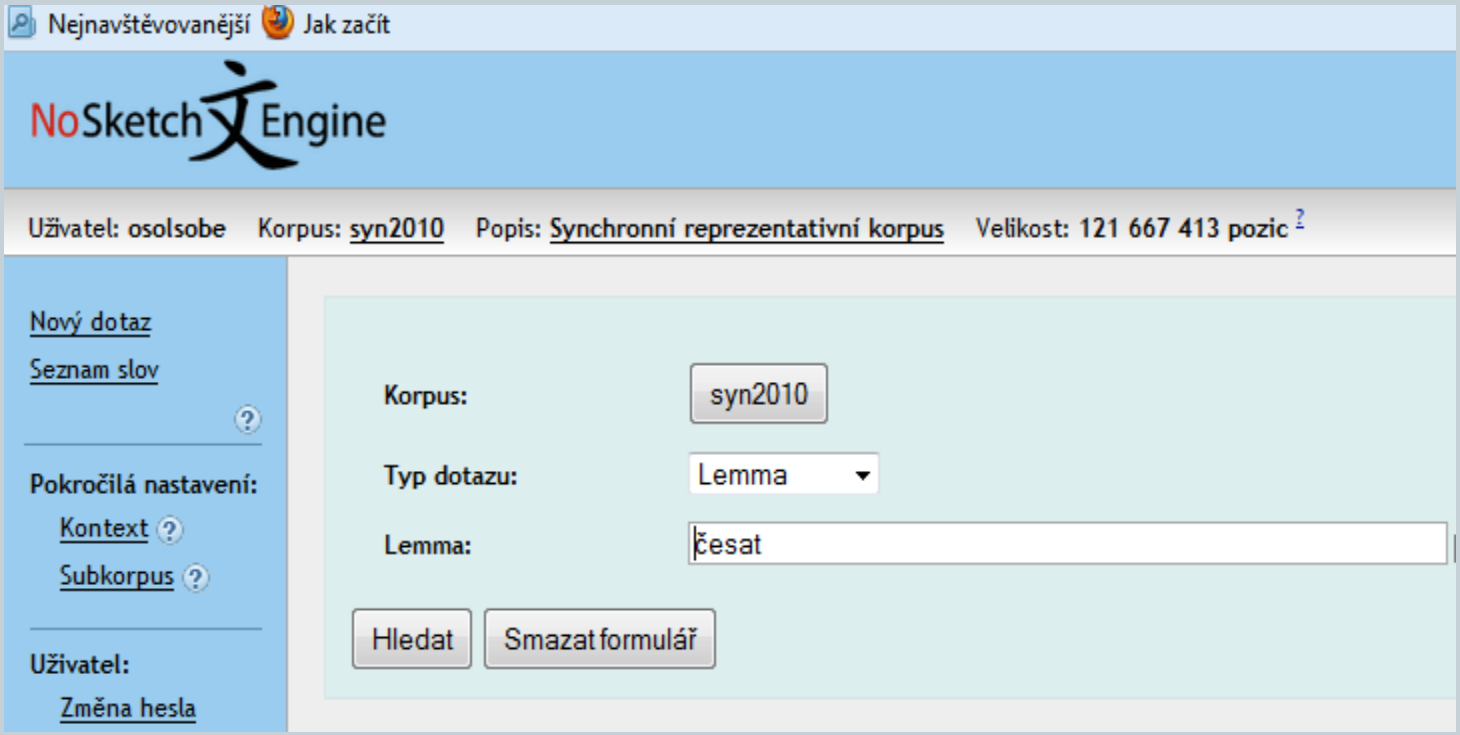
Konkordanční seznam ve formě **KWIC** (Key Word In Context)

- Vyhledá všechny kontexty výskytu hledaného slovního tvaru



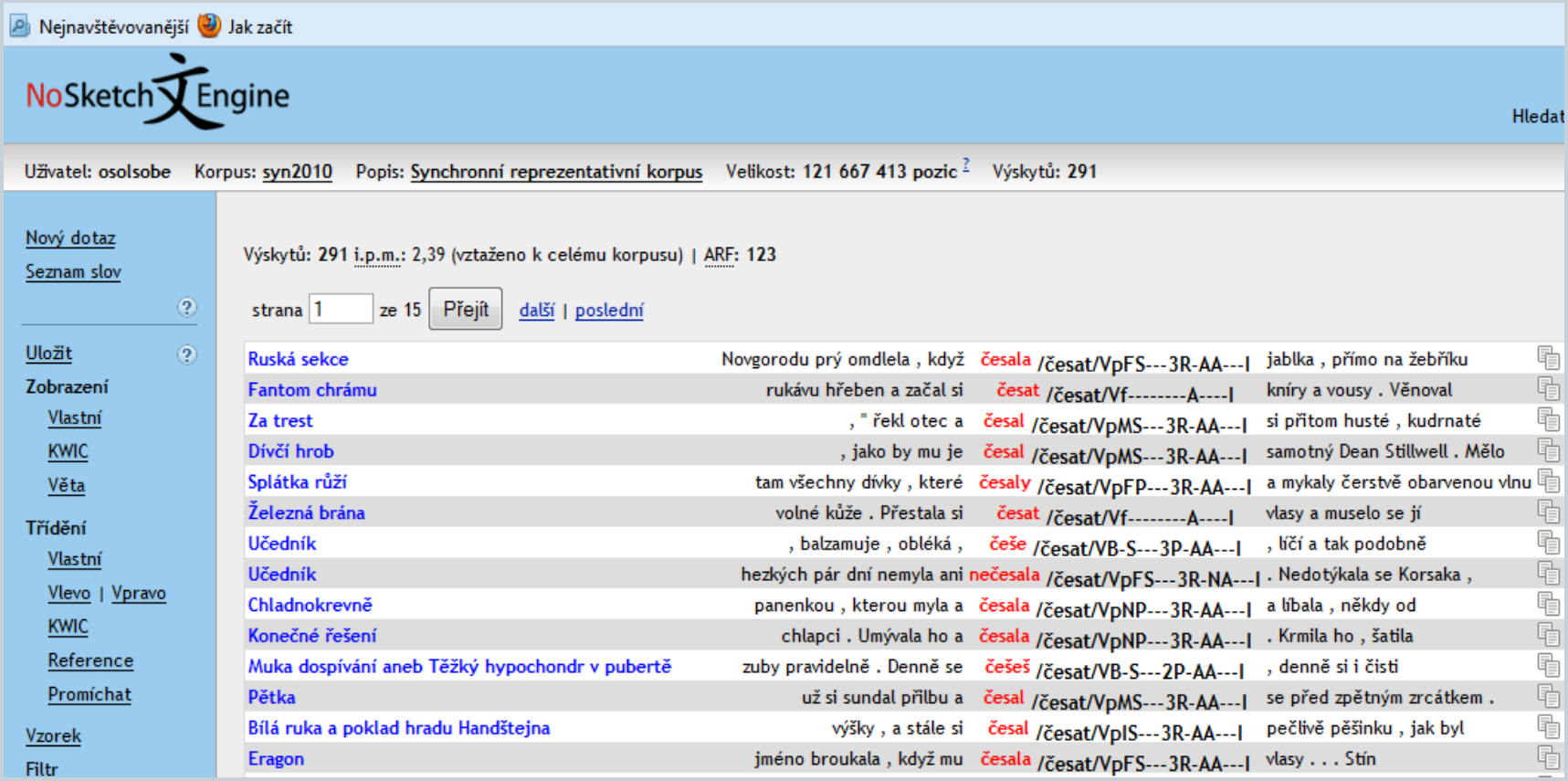
Lemma – základní tvar – systémové slovo

- Lemma (***česat***)



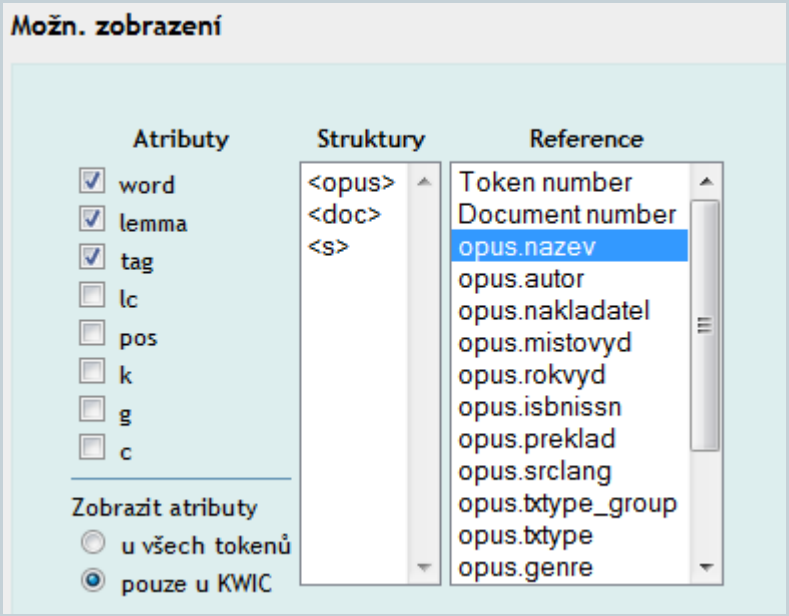
Konkordanční seznam ve formě **KWIC** (Key Word In Context)

- Výskytů tvarů (word) hledaného lemmatu je 291



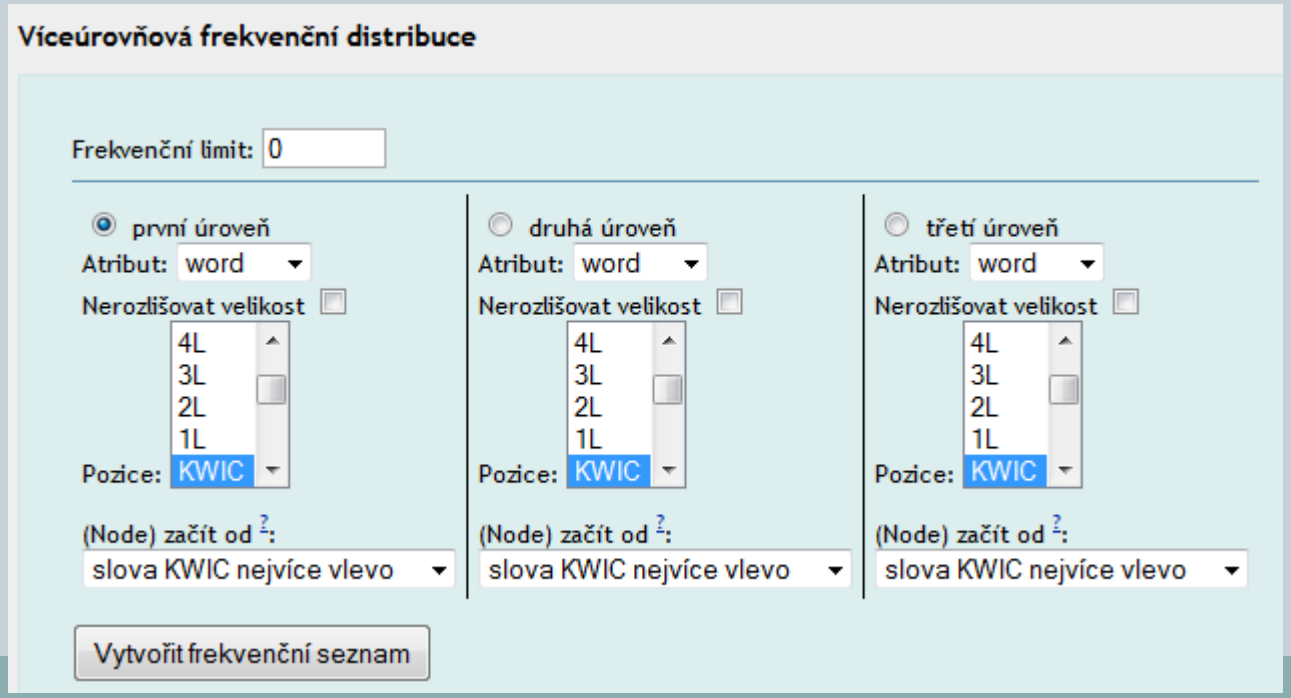
Definování zobrazení hodnot lemmatu, tagu a díla pro KWIC

- Zobrazení



Možnosti zjištění frekvenční distribuce

- Frekvenční distribuce



Frekvenční distribuce slovních tvarů

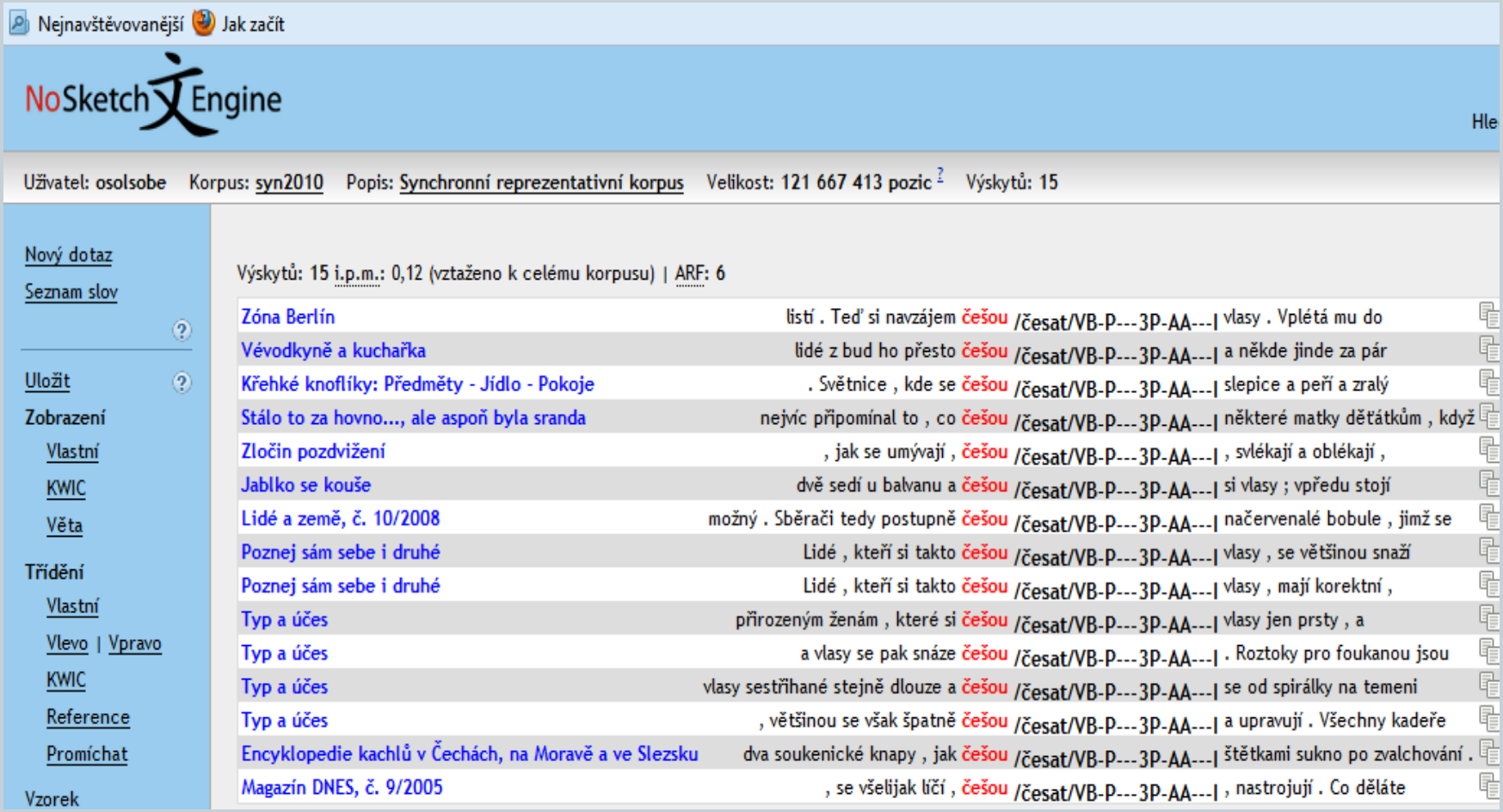
Frekvenční distribuce

Frekvenční limit: 0 Nastavit limit

Celkem: 40 (1 str.)

		word	Frekvence
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
Výskytů: 15
Nový dotaz
Seznam slov
Uložit
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 . Teď si navzájem češou /česat/VB-P---3P-AA---I vlasy . Vplétá mu do
Vévodkyně a kuchařka lidé z bud ho přesto češou /česat/VB-P---3P-AA---I a někde jinde za pár
Křehké knoflíky: Předměty - Jídlo - Pokoje . Světnice , kde se češou /česat/VB-P---3P-AA---I slepice a peří a zralý
Stálo to za hovno..., ale aspoň byla sranda nejvíc připomínal to , co češou /česat/VB-P---3P-AA---I některé matky děťátkům , když
Zločin pozdvižení , jak se umývají , češou /česat/VB-P---3P-AA---I , svlékají a oblékají ,
Jablko se kouše dvě sedí u balvanu a češou /česat/VB-P---3P-AA---I si vlasy ; vpředu stojí
Lidé a země, č. 10/2008 možný . Sběrači tedy postupně češou /česat/VB-P---3P-AA---I načervenalé bobule , jimž se
Poznej sám sebe i druhé Lidé , kteří si takto češou /česat/VB-P---3P-AA---I vlasy , se většinou snaží
Poznej sám sebe i druhé Lidé , kteří si takto češou /česat/VB-P---3P-AA---I vlasy , mají korektní ,
Typ a účes přirozeným ženám , které si češou /česat/VB-P---3P-AA---I vlasy jen prsty , a
Typ a účes a vlasy se pak snáze češou /česat/VB-P---3P-AA---I . Roztoky pro foukanou jsou
Typ a účes vlasy sestřihané stejně dlouze a češou /česat/VB-P---3P-AA---I se od spirálky na temeni
Typ a účes , většinou se však špatně češou /česat/VB-P---3P-AA---I a upravují . Všechny kadeře
Encyklopedie kachlů v Čechách, na Moravě a ve Slezsku dva soukenické knapy , jak češou /česat/VB-P---3P-AA---I štětkami sukno po zvalchování .
Magazín DNES, č. 9/2005 , se všelijak líčí , češou /česat/VB-P---3P-AA---I , nastrojují . Co děláte

Filtry

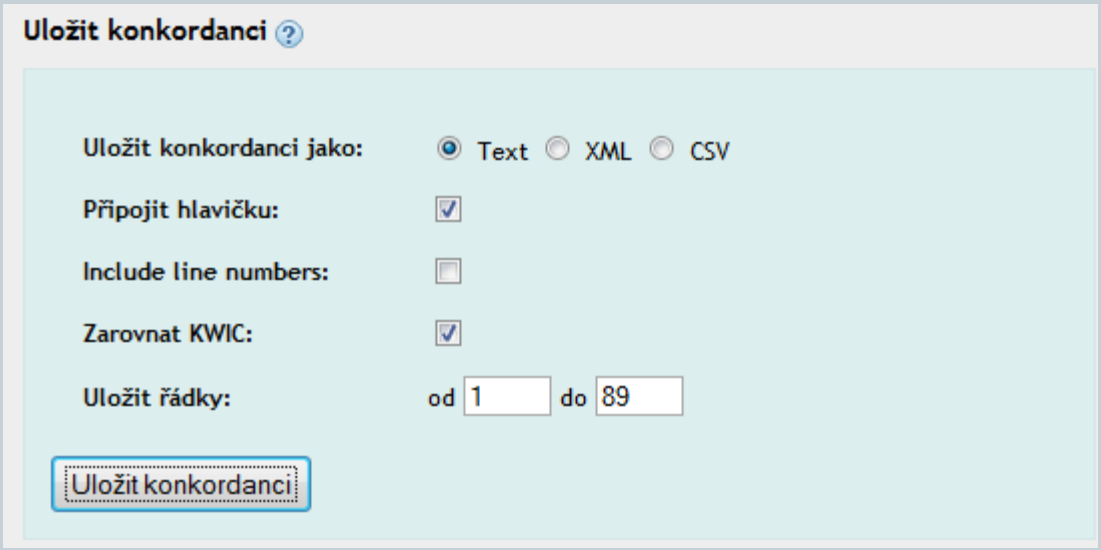
- Na pozici KWIC <0,0> pouze tvary prézentu

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 čisti
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	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	. Většině 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 . Teď si navzájem	češou	/česat/VB-P---3P-AA---I	vlasy . 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 sváželo
Vévodkyně a kuchařka	lidé z bud ho přesto	češou	/česat/VB-P---3P-AA---I	a někde jinde za pár
Chlapectví	že si pomáduje vlasy a	češe	/česat/VB-S---3P-AA---I	se na patku . On
Kainova kniha	plešatí a blonďaté vlasy 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 pěknej	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



Uložení do textového formátu

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

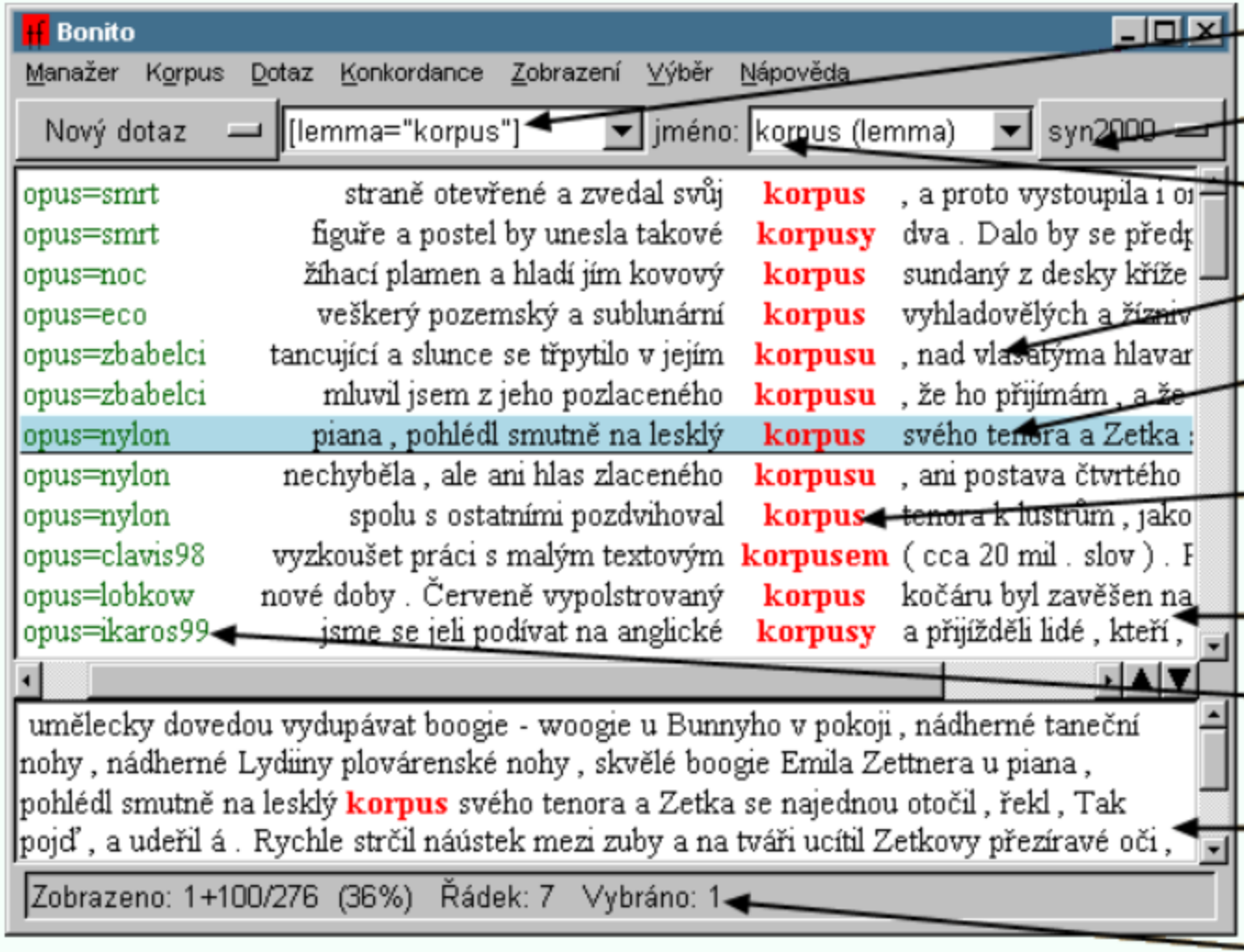
```
# Corpus: syn2010
# Hits: 89
# Relative frequency: 0.73 ((vztaženo k celému korpusu))
# ARF: 39.96
# Query: word,[lemma="česat"] 291
# Positive filter: 0 0 1 [tag="VB.*"] 89

Učedník , balzamuje , obléká , < češe /česat/VB-S---3P-AA---I > , 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 čisti
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 > jeji 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 > vlasy . "
Jsi krásná
Zelená mile 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 > . Většině 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



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 označkován 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.

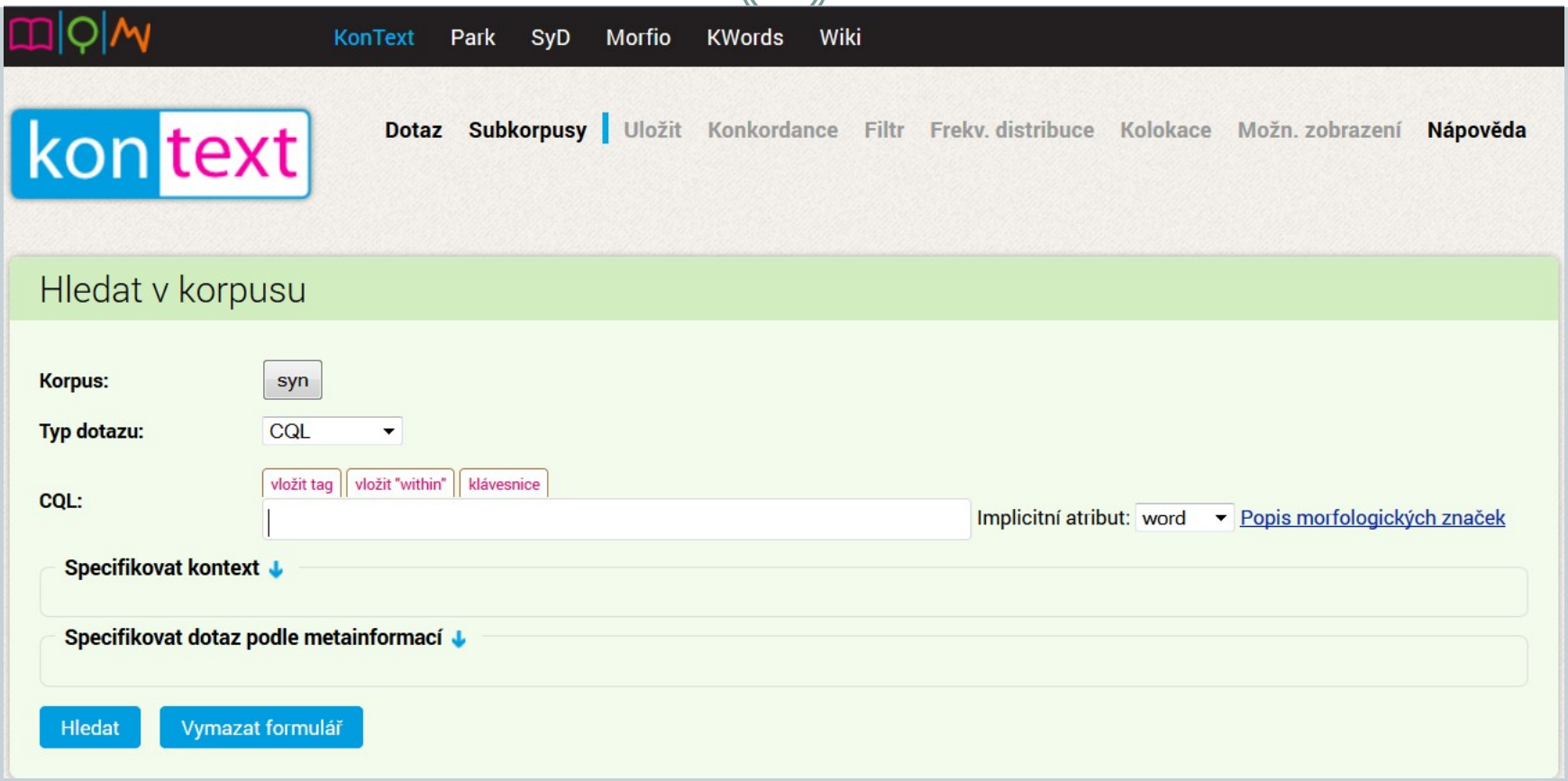
Od ledna 2014 KonText

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



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**