

Úvod do korpusové lingvistiky

5

1

KORPUSOVÉ NÁSTROJE II

Další nástroje přístupné přes rozhraní KonText

2

The image shows a desktop environment with a dark-themed application menu. The menu is open, displaying a list of applications with star icons. The top bar contains icons for a book, a female symbol, and a stylized 'M', followed by text labels: 'Apps', 'WaG', 'KonText', 'Treq', 'Wiki', 'Support', and 'Biblio'. The application list includes 'WaG', 'KonText', 'SyD', 'Morfio', 'KWords', 'Treq', 'Calc', and 'Lists'.

Application	Status
WaG	★
KonText	★
SyD	☆
Morfio	☆
KWords	☆
Treq	★
Calc	☆
Lists	☆

WaG/Word at a Glance/Slovo v kostce

3

- Find a word (*sníh, sucho, válka*)
- Compare words (*snídaně, oběd, večeře/ Německo, Rusko/ chytrý, hloupý*)
- Search in two languages / translate (*host, opilec*)

Treq

4

Treq

DATABÁZE
PŘEKLADOVÝCH
EKVIVALENTŮ

Ver. 2.2

Výchozí jazyk: Čeština

Cílový jazyk: Francouzština

Omezit na: Kolekce: 7

daňová povinnost Hledej

Lemmata Víceslovné Regulární A = a

▲ Frekvence ▼	▲ Procenta ▼	▲ Čeština ▼	▲ Francouzština ▼
15	53.6	daňová povinnost	exigible
4	14.3	daňová povinnost	la taxe exigible
3	10.7	daňová povinnost	taxe exigible
1	3.6	daňová povinnost	exigibilité de
1	3.6	daňová povinnost	assujettissement de
1	3.6	daňová povinnost	accise devient exigible
1	3.6	daňová povinnost	exigibilité la taxe
1	3.6	daňová povinnost	exigibilité
1	3.6	daňová povinnost	assujettissement

Jaký host ?

5

Treq

DATABÁZE
PŘEKLADOVÝCH
EKVIVALENTŮ

Ver. 2.2

Výchozí jazyk: Čeština

Cílový jazyk: Francouzština

Omezit na: Kolekce: 7

.*ý host Hledej

Lemmata Víceslovné Regulární A = a

▲ Frekvence ▼	▲ Procenta ▼	▲ Čeština ▼	▲ Francouzština ▼
14	8.1	stálý host	habitué
4	2.3	čestný host	invité
4	2.3	stálý host	le habitué
3	1.7	stálý host	habitué du
3	1.7	zvaný host	intrus
3	1.7	čestný host	invité de honneur
3	1.7	zvaný host	invité
2	1.2	pozvaný host	invité
2	1.2	vítaný host	le bienvenu
2	1.2	nečekaný host	surpris

Morfio

6

- Hledání n-tic se společnou a odlišnou částí
- Možnost sledovat slovotvornou produktivitu
- Možnost vyhledávat varianty

ý/ota

7

Morfio

Jazyk: čeština ▾

<+ ✖ společný ▾

✖ odlišný ▾

+> Morf. specifikace:

vzor 1:

.+

ý

přídavná jména ▾

A.*

vzor 2:

.+

ota

podstatná jména ▾

N.*

Přidat vzor

Korpus: SYN2005 ▾

Frekvence vyšší než: 0

Hledat: lemmata ▾

Vyhodnotit: lemmata ▾

A = a

▶ Alternace

Hledat

Nové zadání

Odkaz na toto zadání: <http://morfio.korpus.cz/9pLHBjCF>

Nápověda

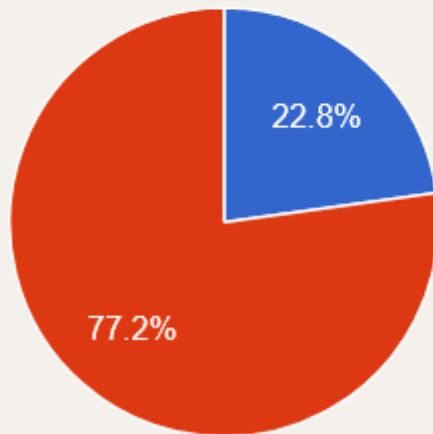
Přegenerované výsledky

8

Souhrn Výpis Produktivita vzor 1 vzor 2 Souhrn Výpis Produktivita vzor 1

Souhrn	celkem tokenů	typů	v modelu tokenů	typů
vzor 1	8471877	74252	655690	89
vzor 2	115198	390	84722	89

Odhad úplnosti daného modelu



● v modelu
● nepokryto

	vzor 1 ▲▼ (fq ▲▼)	vzor 2 ▲▼ (fq ▲▼)
1	hodný (5887)	hodnota (30888)
2	teplý (8320)	teplota (10804)
3	bý (3)	bota (7958)
4	jistý (30680)	jistota (7758)
5	samý (39)	samota (3726)
6	jedný (3)	jednota (3277)
7	nejistý (2)	nejistota (2822)
8	čistý (14480)	čistota (2195)
9	temný (6584)	temnota (1900)
10	ochý (1)	ochota (1687)
11	ný (32)	nota (1512)
12	hustý (5090)	hustota (1504)
13	prázdný (10074)	prázdnota (1476)
14	dobrý (103080)	dobrota (1170)
15	řý (1)	rota (1160)
16	nahý (4383)	nahota (532)
17	lhý (2)	lhota (527)
18	prostý (6758)	prostota (438)
19	nový (138367)	novota (390)
20	slepý (3573)	slepotá (385)

Rychle je patrný rozdíl mezi morfem a pouhým řetězcem (homonymie)

9

Souhrn	Výpis	Produktivita	vzor 1	vzor																																																		
<input checked="" type="checkbox"/> skrýt formy s podlimitní frekvencí		Výsledky analýzy zvoleném korpusu zvýrazněné barev																																																				
<table border="1"><thead><tr><th>abc/cba ▲▼</th><th>fq ▲▼</th></tr></thead><tbody><tr><td>hodnota</td><td>30888</td></tr><tr><td>sobota</td><td>12048</td></tr><tr><td>teplota</td><td>10804</td></tr><tr><td>bota</td><td>7958</td></tr><tr><td>jistota</td><td>7758</td></tr><tr><td>hmota</td><td>6702</td></tr><tr><td>samota</td><td>3726</td></tr><tr><td>jednota</td><td>3277</td></tr><tr><td>nejjistota</td><td>2822</td></tr><tr><td>čistota</td><td>2195</td></tr><tr><td>porota</td><td>2116</td></tr><tr><td>temnota</td><td>1900</td></tr><tr><td>ochota</td><td>1687</td></tr><tr><td>nota</td><td>1512</td></tr><tr><td>hustota</td><td>1504</td></tr><tr><td>prázdnota</td><td>1476</td></tr><tr><td>dobrota</td><td>1170</td></tr><tr><td>rota</td><td>1160</td></tr><tr><td>pěchota</td><td>1033</td></tr><tr><td>toyota</td><td>956</td></tr><tr><td>nečistota</td><td>821</td></tr><tr><td>kapota</td><td>658</td></tr><tr><td>anekdota</td><td>608</td></tr><tr><td>neochota</td><td>588</td></tr></tbody></table>		abc/cba ▲▼	fq ▲▼	hodnota	30888	sobota	12048	teplota	10804	bota	7958	jistota	7758	hmota	6702	samota	3726	jednota	3277	nejjistota	2822	čistota	2195	porota	2116	temnota	1900	ochota	1687	nota	1512	hustota	1504	prázdnota	1476	dobrota	1170	rota	1160	pěchota	1033	toyota	956	nečistota	821	kapota	658	anekdota	608	neochota	588	Tabulky vyhodnoceno zahrnuto nebylo (Tabulku je možné retrográdní (tedy Řádky se v takové písmeno od začátku dané skupiny, a to		
abc/cba ▲▼	fq ▲▼																																																					
hodnota	30888																																																					
sobota	12048																																																					
teplota	10804																																																					
bota	7958																																																					
jistota	7758																																																					
hmota	6702																																																					
samota	3726																																																					
jednota	3277																																																					
nejjistota	2822																																																					
čistota	2195																																																					
porota	2116																																																					
temnota	1900																																																					
ochota	1687																																																					
nota	1512																																																					
hustota	1504																																																					
prázdnota	1476																																																					
dobrota	1170																																																					
rota	1160																																																					
pěchota	1033																																																					
toyota	956																																																					
nečistota	821																																																					
kapota	658																																																					
anekdota	608																																																					
neochota	588																																																					

Varianty

10

Morfio

Jazyk: čeština ▾

<+ ✖ společný ▾ ✖ odlišný ▾ +> Morf. specifikace:

vzor 1: .+ ▾ i ▾ vlastní tag > ▾ NNMP1.*

vzor 2: .+ ▾ ové ▾ vlastní tag > ▾ NNMP1.*

Přidat vzor

Korpus: SYN2005 ▾ Frekvence vyšší než: 0 Hledat: tvary ▾ Vyhodnotit: tvary ▾

A = a

▶ Alternace

Hledat Nové zadání Odkaz na toto zadání: <http://morfio.korpus.cz/I96TOKdU> Nápověda

Podle frekvence vzor1/vzor2

11

	vzor 1 ▲▼ (fq ▲▼)	vzor 2 ▲▼ (fq ▲▼)
1	muži (11980)	mužové (471)
2	zástupci (3608)	zástupcové (8)
3	hasiči (2335)	hasičové (1)
4	psi (2272)	psové (15)
5	páni (1778)	pánové (2286)
6	francouzi (1327)	francouzové (18)
7	čtenáři (1250)	čtenářové (1)
8	hoši (1226)	hošové (1)
9	tvůrci (1154)	tvůrcové (14)
10	zloději (1067)	zlodějové (4)
11	soudci (954)	soudcové (104)
12	jedinci (942)	jedincové (1)
13	odpůrci (846)	odpůrcové (2)
14	milenci (786)	milencové (1)
15	samci (537)	samcové (1)
16	strážci (521)	strážcové (51)
17	domorodci (513)	domorodcové (1)
18	zastánci (506)	zastáncové (1)
19	vůdci (505)	vůdcové (167)
20	šamani (463)	šamanové (9)

	vzor 1 ▲▼ (fq ▲▼)	vzor 2 ▲▼ (fq ▲▼)
1	členi (4)	členové (6519)
2	páni (1778)	pánové (2286)
3	rusi (10)	rusové (1912)
4	briti (2)	britové (1603)
5	syni (4)	synové (1152)
6	otci (289)	otcové (675)
7	romi (1)	romové (662)
8	králi (78)	králové (518)
9	vítězi (86)	vítězové (513)
10	muži (11980)	mužové (471)
11	biskupi (3)	biskupové (456)
12	arabi (8)	arabové (383)
13	švédi (17)	švédové (355)
14	dáni (3)	dánové (295)
15	elfi (1)	elfové (258)
16	indi (5)	indové (255)
17	vůdci (505)	vůdcové (167)
18	posli (11)	poslové (147)
19	dábli (11)	dáblové (127)
20	lumpi (5)	lumpové (117)



Diachronní

Synchronní

Nový dotaz

[1] nemůžu

[2] nemohu



Porovnat varianty



a=A



lemma

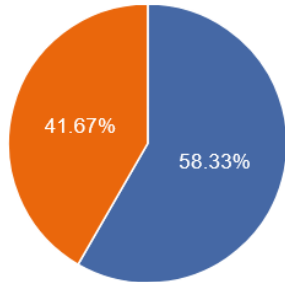
Zobrazit adresu dotazu



Sdílet

Psaný jazyk

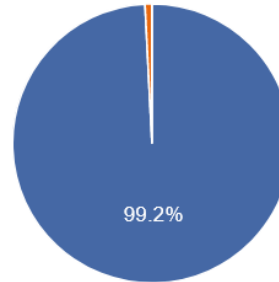
■ [1] nemůžu
■ [2] nemohu



Uložit jako ▾

Mluvený jazyk

■ [1] nemůžu
■ [2] nemohu



Uložit jako ▾

Celkové údaje p
vzhledem k ve
Přesné údaje o p
oblast grafu.

Údaj *Nedostatek*
variant v daném

Zadání dotazu a
být znovu vyvo
stránky. Odkaz je

háziš/hážeš

13

SyD

Diachronní

Synchronní

Nový dotaz

[1] .*[^c]ház(i[mš]|i[mt]e|íejí) [2] .*ház([eu]|ou|eš|e[mt]e) +

Porovnat varianty



a=A



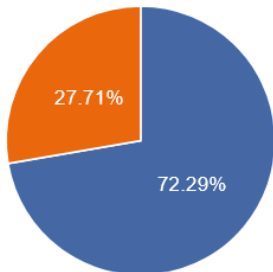
lemma

Zobrazit adresu dotazu



Sdílet

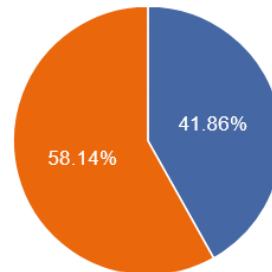
Písaný jazyk



■ [1] .*[^c]ház(i[mš]|i[mt]e|íejí)
■ [2] .*ház([eu]|ou|eš|e[mt]e)

Uložit jako ▾

Mluvený jazyk



■ [1] .*[^c]ház(i[mš]|i[mt]e|íejí)
■ [2] .*ház([eu]|ou|eš|e[mt]e)

Uložit jako ▾

Celkové údaje pro psaný jazyk vzhledem k velké oblasti. Přesné údaje o poměrech v dané oblasti grafu.

Údaj *Nedostatečná data* znamená, že některé varianty v daném (sub)kontextu nejsou zastoupeny.

Zadání dotazu a jeho výsledky mohou být znovu vyvolány pomocí odkazu na stránce. Odkaz je možný pouze pro psaný jazyk.

Ukázka otázek v testu

14

- Lze ve všech nástrojích, které nabízí rozhraní Kontext, pracovat se zadáním dotazu pomocí regulárních výrazů?
- Vyber si jeden z nástrojů nabízených rozhraním KonText, který jsi vyzkoušel, a popiš úkol, který tě zaujal.
- Které z nástrojů nabízených rozhraním KonText nabízejí data čerpaná z paralelního korpusu?
- Jaká jsou omezení funkcí, které porovnávají data z psaných a mluvených korpusů?