PLIN033_2

MORFIO

Aplikace k odhadování rozsahu a produktivity slovotvorných modelů v češtině na základě korpusových dat

Co dnes chceme?

- představit aplikaci *MORFIO*
- ukázat na konkrétním příkladu postup práce při extrakci podkladů pro lingvistickou analýzu slovotvorných formací automaticky získaných prostřednictvím aplikace *MORFIO*
- zadat úkoly na příště

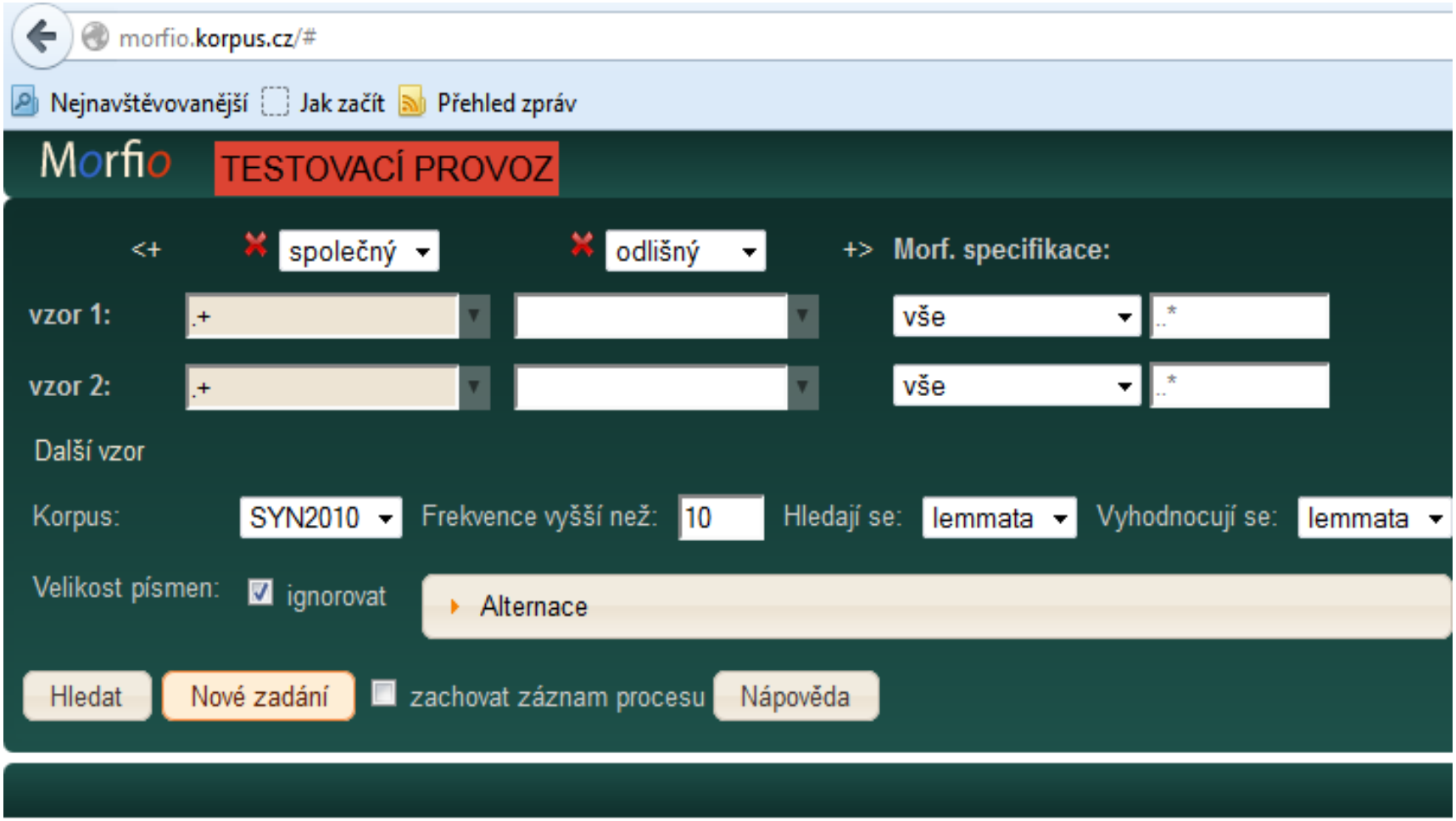
Co je to *MORFIO ?*

- Aplikace *Morfio* slouží k odhadování rozsahu a produktivity slovotvorných modelů v češtině na základě korpusových dat.
- slovotvorný vztah - vytvářen 1) formální shodou/podobností v určitých částech slova, tzv. báze (např. *dřev-* je část společná pro slova *dřevo* i *dřevěný*) a 2) formálními odlišnostmi v částech specifických, tzv. formantech (morfy *-o* a *-ěný* v předchozím příkladu).
- Cílem aplikace je najít všechny dvojice, resp. trojice nebo čtveřice, jednotek v korpusu, které se shodují v bázi a liší se pouze specifikovanými formanty.
- Výstupem aplikace *Morfio* není a nemůže být lingvisticky bezchybný výstup, spíš se jedná o pomůcku, která množství dat dokáže pro lingvistické účely předzpracovat.

MORFIO

- http://morfio.korpus.cz/
- http://ucnk.ff.cuni.cz/bonito/znacky.php
- http://wiki.korpus.cz/doku.php/pojmy:morfologicka_analyza
- http://wiki.korpus.cz/doku.php/pojmy:tag

Jak vypadá přístup



Volba korpusu

- SYN2010
- SYN2005
- SYN2015
- Araneum CS Minus

https://wiki.korpus.cz/doku.php/cnk:uvod

Přehled dostupných korpusů

Korpusy psaného jazyka (synchronní)					
korpus	**velikost (počet slov)**	**lemmatizace**	**morfologické značky**	**rok zveřejnění**[1)]	**charakteristika korpusu**
Obecné korpusy					
SYN (verze 8)	4,5 mld.	√	√	2010-2019	verzovaný korpus, spojující synchronní psané korpusy řady SYN a další, dosud nezveřejněné texty
SYN2020	100 mil.	√	√	2020	referenční reprezentativní korpus, převažují texty z let 2015–2019
SYN2015	100 mil.	√	√	2015	referenční reprezentativní korpus, převažují texty z let 2010–2014, s novou klasifikací textů
SYN2013PUB	935 mil.	√	√	2013	referenční korpus publicistických textů z let 2005–2009
SYN2010	100 mil.	√	√	2010	referenční reprezentativní korpus, převažují texty z let 2005–2009
SYN2009PUB	700 mil.	√	√	2010	referenční korpus publicistických textů z let 1995–2007
SYN2006PUB	300 mil.	√	√	2006	referenční korpus publicistických textů z let 1989–2004
SYN2005	100 mil.	√	√	2005	referenční reprezentativní korpus, převažují texty z let 2000–2004
SYN2000	100 mil.	√	√	2000	referenční reprezentativní korpus, převažují texty z let 1990–1999

Zadání dotazu

- regulární výrazy
- morfologické značky (viz výše)
- další možnosti zobecnění (rozlišení samohlásek, souhlásek, ... Viz Nápověda)
- volby různých typů alternací

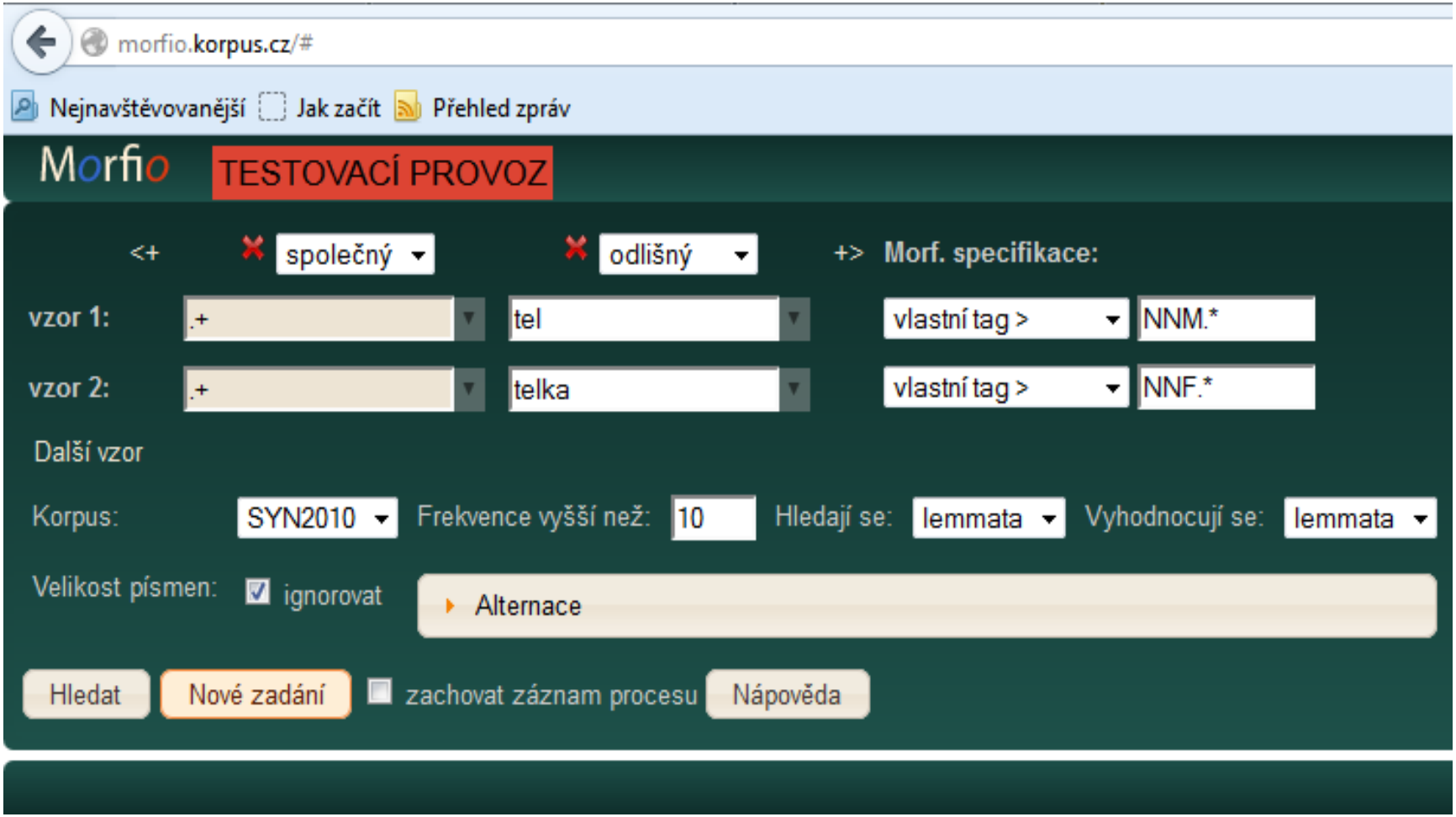
Co chceme?

- Kolik a která substantiva typu *učitel* mají ženský protějšek (přechýlování/moce) tvořený příponou *-ka*
- Jaké další dvojice tohoto typu známe?
- Jak je můžeme popsat?
- Jaké mají formální vlastnosti?

-tel / -telka

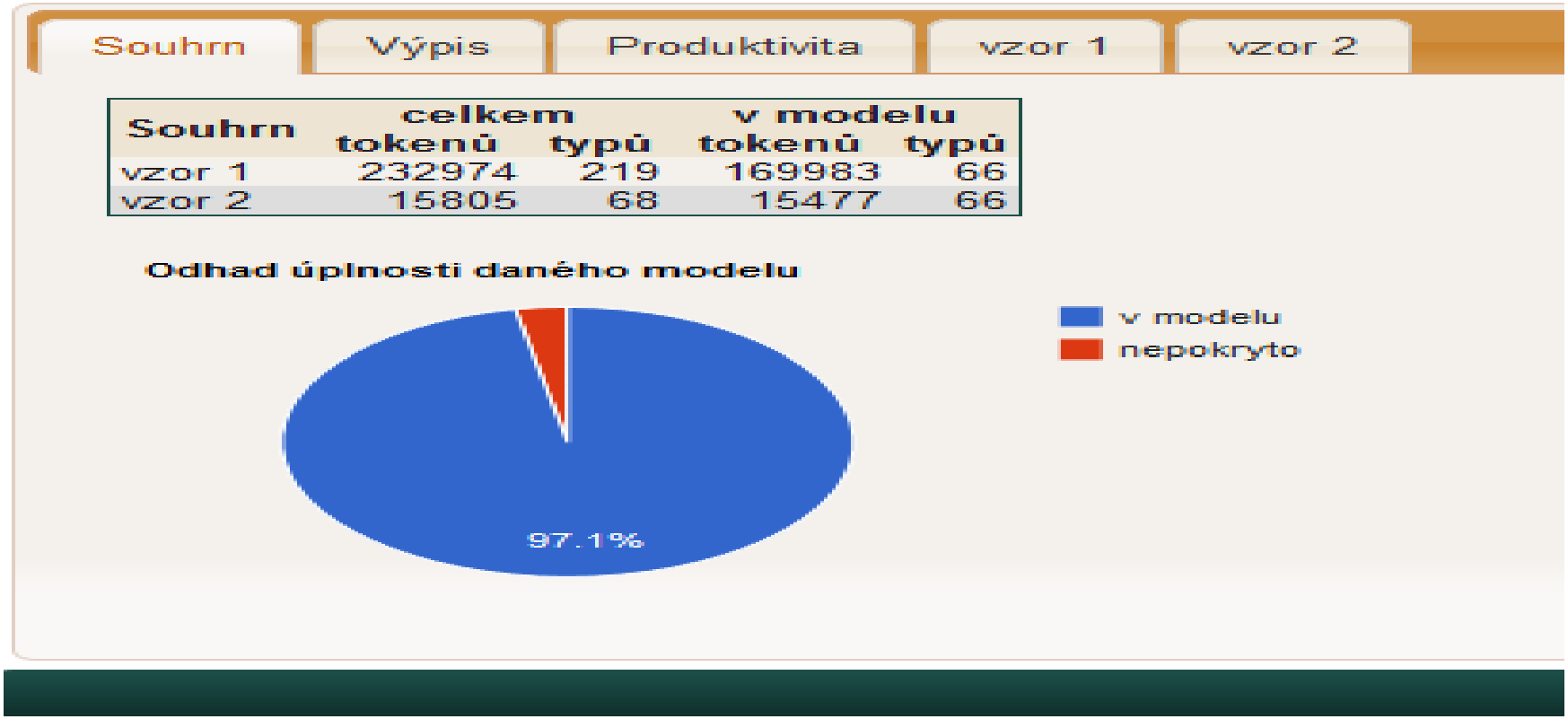
- Substantivum maskulinum životné
- Lemma končí na *tel*
- Substantivum femininum
- Lemma končí na *telka*

Vyplnění formuláře



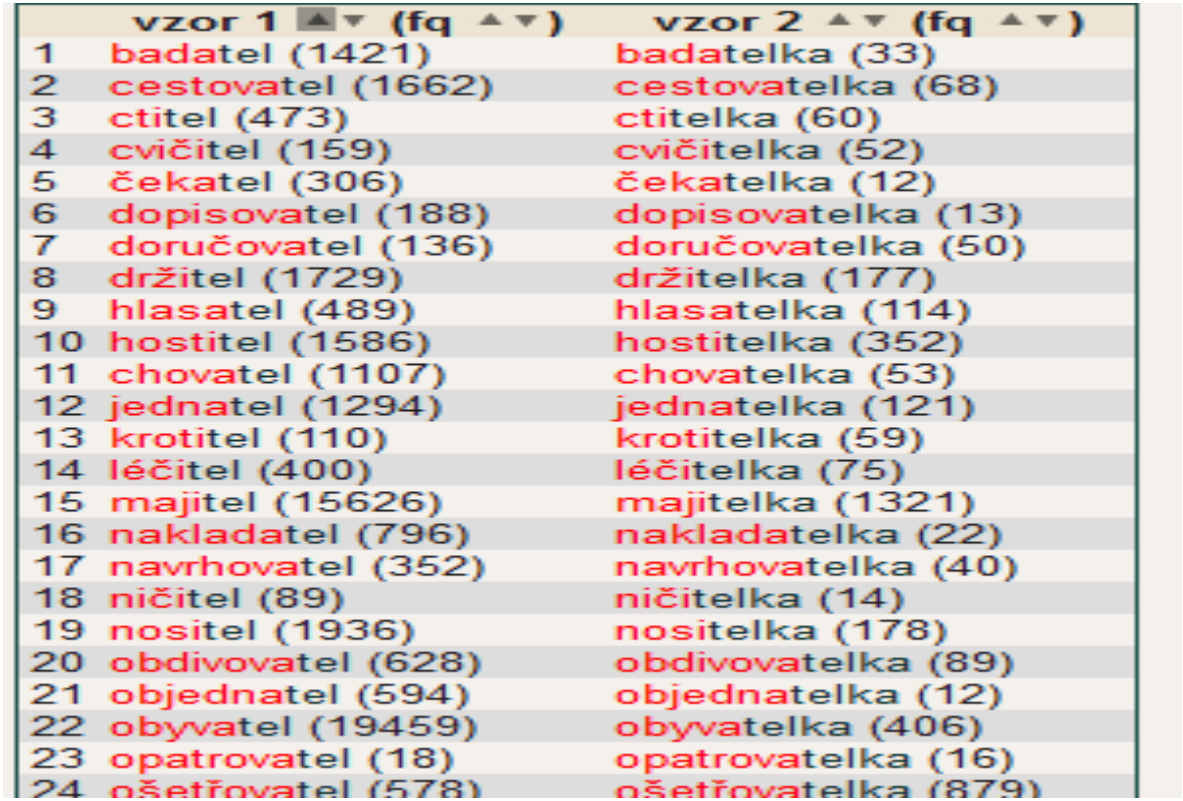
SOUHRN

V záložce **souhrn** jsou uvedeny počty typů s nadlimitní frekvencí a součet jejich výskytů. Jedna sada údajů (sloupec "celkem") se vždy týká vzoru samotného (chápaného izolovaně), druhá sada (sloupec "v modelu") pak odkazuje k těm jednotkám příslušejícím ke vzoru, které zároveň patří do analyzovaného slovotvorného modelu, tj. slova, která mají k sobě odvozeninu identifikovanou v rámci druhého vzoru.



Výpis

V tabulce jsou uvedeny všechny doklady ze všech vzorů, které vstupují do zadaného modelu. Červená část slov označuje společnou bázi (ta se může lišit pouze v případě aplikace alternací). V závorkách uvedený údaj představuje celkovou frekvenci jednotky ve zvoleném korpusu. Tabulku je možné přetřídit podle libovolného sloupce a to jak abecedně, tak frekvenčně pomocí šipek v záhlaví tabulky. Každé slovo zároveň funguje jako odkaz směřující k ukázce konkordancí ve zvoleném korpusu.



	vzor 1 (fq)	vzor 2 (fq)
1	badatel (1421)	badatelka (33)
2	cestovatel (1662)	cestovatelka (68)
3	ctitel (473)	ctitelka (60)
4	cvičitel (159)	cvičitelka (52)
5	čekatel (306)	čekatelka (12)
6	dopisovatel (188)	dopisovatelka (13)
7	doručovatel (136)	doručovatelka (50)
8	držitel (1729)	držitelka (177)
9	hlasatel (489)	hlasatelka (114)
10	hostitel (1586)	hostitelka (352)
11	chovatel (1107)	chovatelka (53)
12	jednatel (1294)	jednatelka (121)
13	krotitel (110)	krotitelka (59)
14	léčitel (400)	léčitelka (75)
15	majitel (15626)	majitelka (1321)
16	nakladatel (796)	nakladatelka (22)
17	navrhovatel (352)	navrhovatelka (40)
18	ničitel (89)	ničitelka (14)
19	nositel (1936)	nositelka (178)
20	obdivovatel (628)	obdivovatelka (89)
21	objednatel (594)	objednatelka (12)
22	obyvatel (19459)	obyvatelka (406)
23	opatrovatel (18)	opatrovatelka (16)
24	ošetřovatel (578)	ošetřovatelka (879)

Výpis lze

- seřadit podle frekvence 1. nebo 2. členu dvojice

	vzor 1 (fq)	vzor 2 (fq)
1	ředitel (24415)	ředitelka (4025)
2	obyvatel (19459)	obyvatelka (406)
3	majitel (15626)	majitelka (1321)
4	učitel (10943)	učitelka (3293)
5	podnikatel (9283)	podnikatelka (334)
6	uživatel (8373)	uživatelka (22)
7	spisovatel (7544)	spisovatelka (837)
8	představitel (6912)	představitelka (384)
9	velitel (6557)	velitelka (129)
10	zaměstnavatel (5810)	zaměstnavatelka (54)
11	pachatel (5612)	pachatelka (54)
12	zastupitel (5086)	zastupitelka (189)
13	pořadatel (3701)	pořadatelka (59)
14	spotřebitel (2839)	spotřebitelka (17)
15	provozovatel (2763)	provozovatelka (55)
16	zakladatel (2718)	zakladatelka (174)
17	skladatel (2411)	skladatelka (69)
18	vyšetřovatel (2394)	vyšetřovatelka (95)
19	žadatel (2171)	žadatelka (23)
20	pozorovatel (2129)	pozorovatelka (21)

Výpis lze

- seřadit abecedně podle 1. nebo 2. členu dvojice

	vzor 1 (fq)	vzor 2 (fq)
1	badatel (1421)	badatelka (33)
2	cestovatel (1662)	cestovatelka (68)
3	ctitel (473)	ctitelka (60)
4	cvičitel (159)	cvičitelka (52)
5	čekatel (306)	čekatelka (12)
6	dopisovatel (188)	dopisovatelka (13)
7	doručovatel (136)	doručovatelka (50)
8	držitel (1729)	držitelka (177)
9	hlasatel (489)	hlasatelka (114)
10	hostitel (1586)	hostitelka (352)
11	chovatel (1107)	chovatelka (53)
12	jednatel (1294)	jednatelka (121)
13	krotitel (110)	krotitelka (59)
14	léčitel (400)	léčitelka (75)
15	majitel (15626)	majitelka (1321)
16	nakladatel (796)	nakladatelka (22)
17	navrhovatel (352)	navrhovatelka (40)
18	ničitel (89)	ničitelka (14)
19	nositel (1936)	nositelka (178)
20	obdivovatel (628)	obdivovatelka (89)

Kliknutím na lemma lze získat konkordanci v příslušném korpusu

Příklady konkordancí pro "[lemma="(?i)badatel" & tag="NNM.*"]" z korpusu syn2010

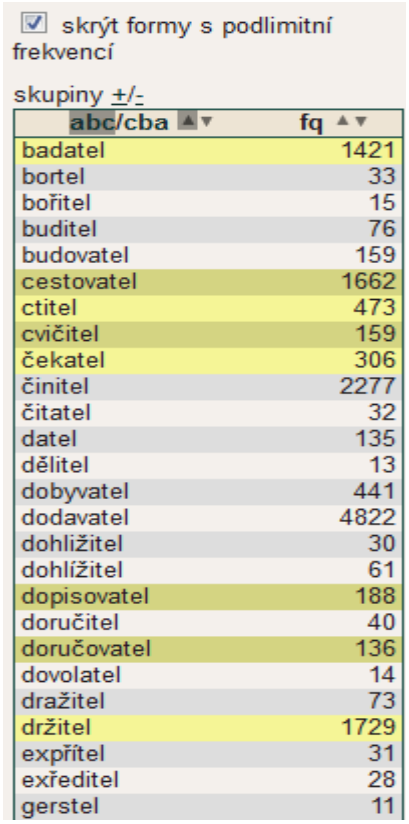
Fuks, Ladislav: Vévodkyně a kuchařka	společníci - kapitán navíc jistý	badatel	, cestovatel a znalec Afriky
Strachey, Lytton: Alžběta a Essex	úzkostném napětí a některé moderní	badatele	vedla k domněnce , že
Y: Austrálie	Průkopníci a badatelé Mnoho evropských	badatelů	, včetně Edwarda Eyra a
Gieselmann, David: Pan Kolpert	takovou naprosto banální odpověď :	badatel	v oblasti chaosu - tedy
Mácha, Karel: Integrální antropolo…	snad zdát tato formulace konkrétnímu	badateli	v některé přírodovědní disciplíně příliš
Vacková, Jarmila: Odpovědi obrazů - Mi…	Malerei (1925) německý	badatel	M. J. Friedländer s tím
Gurevič, Aron Jakovle…: Nebe, peklo, svět - …	s tvou služebnicí ? "	Badatelé	spatřují v rituálu " snižování
Gurevič, Aron Jakovle…: Nebe, peklo, svět - …	Evropě 13 . století odhalují	badatelé	na několik stovek , vedlo
Brunschvicg, Léon: Evropský duch	: " Cítím se být	badatelem	. Žízním po úplném poznání
Bauman, Zygmunt: Úvahy o postmoderní době	městské kultury ; pro mnoho	badatelů	se stal zevloun hlavním symbolem
Potůček, Martin (ed.): Manuál prognostickýc…	souvislosti , které nejsou zřejmé	badatelům	v individuálních výzkumných oblastech .
Keller, Jan: Teorie modernizace	sponzorů ztrácejí univerzitní učitelé a	badatelé	status nezávislých odborníků . Část
Y: Ikarie, č. 4/2005	Cristea , občan Ukrajiny a	badatel	. Měl jednu bodnou ránu
Studnička, Miloslav: Masožravé rostliny	darlingtonií a pavoukem zmiňují různí	badatelé	, včetně historických zpráv zmíněné
Benda, Ivo A. (et al.): VI. rozhovory s pouč…	Uvědomte si proto , pozemští	badatelé	, že to , co
Y: Technik, č. 7/2005	. let dvacátého století se	badatelům	nepodařilo dostat se při snižování
Koukolík, František -…: Šimpanz a vesmír	říkají stochastickými - vlivy .	Badatelé	zkoumající dlouověkost drosofil se samozřejmě
Štoll, Ivan: Podivuhodné přírodní úkazy	1597) , holandský polární	badatel	, když v roce 1596
Wolf, Josef: Kdo je kdo není v hn…	se vyvíjela od původního kroužku	Badatelů	Bible v sedmdesátých letech 19
Y: S tebou mě baví život, č. 2/2008	F. X. Šalda . Literární	badatel	Radko Pytlík však shrnul spisovatelovo
Kalina, Pavel - Koťát…: Praha 1310-1437	jižní pilíř unikal pozornosti novodobých	badatelů	. Je to patrně dáno
Y: Právo, 3. 9. 2005	z pozdní doby bronzové našel	badatel	spolupracující s mladoboleslavským muzeem v
Y: Lidové noviny, 17. 7. 2006	je nezískal vůbec . Cambridgeští	badatelé	řadu let zkoumali třináct skupin
Y: Mladá fronta DNES, 25. 4. 2007	poškozené spisy neměly být přístupné	badatelům	. " Můžeme ho ukázat
Y: Sedmička, č. 18/2009	, spoustu zajímavých informací však	badatelé	objeví i v hradeckém Okresním

Vzor 1

- Výsledky analýzy jednotlivých vzorů jako samostatných dotazů jsou prezentovány ve formě tabulky jednotek (slovních tvarů nebo lemmat) spolu s jejich frekvencemi ve zvoleném korpusu. Tabulku je možné doplnit i o jednotky, které v modelu nebyly brány v potaz, protože jejich frekvence byla nižší než uživatelem stanovený limit. Údaje zvýrazněné barevným pozadím se účastní slovotvorného modelu (tj. existuje k nim v druhém vzoru protějšek se stejnou bází, lišící se pouze formanty).

tel (vzor 1)

- Žlutě jsou ty členy, k nimž byl nalezen vzor 2



☑ skrýt formy s podlimitní frekvencí

skupiny +/-

abc/cba	fq
badatel	1421
bortel	33
bořitel	15
buditel	76
budovatel	159
cestovatel	1662
ctitel	473
cvičitel	159
čekatel	306
činitel	2277
čitatel	32
datel	135
dělitel	13
dobyvatel	441
dodavatel	4822
dohližitel	30
dohlížitel	61
dopisovatel	188
doručitel	40
doručovatel	136
dovolatel	14
dražitel	73
držitel	1729
expřítel	31
exředitel	28
gerstel	11

telka (vzor2)

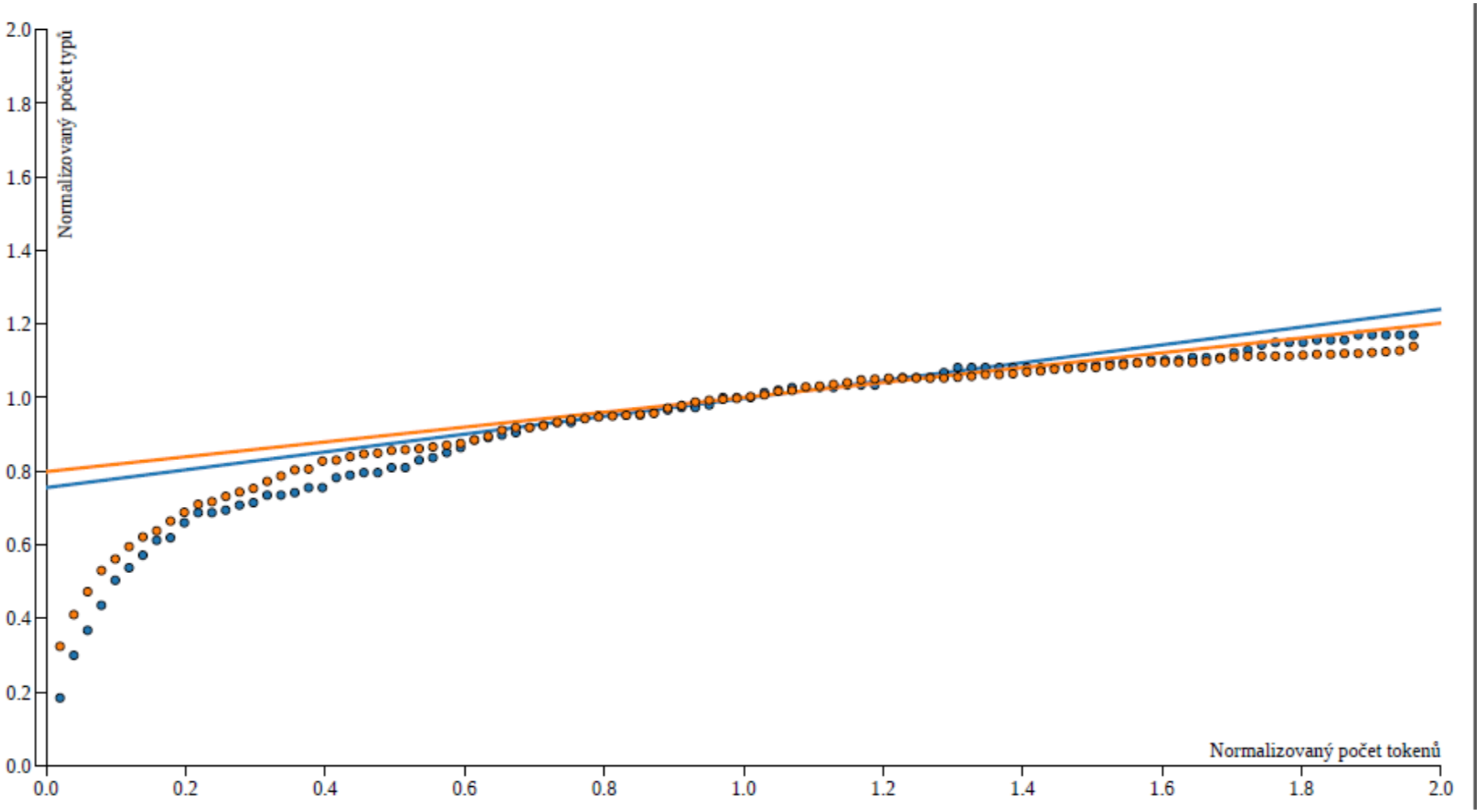
- Žlutě jsou ty členy, k nimž byl nalezen vzor 1

☑ skrýt formy s podlimitní frekvencí

skupiny +/-

abc/cba	fq
badatelka	33
cestovatelka	68
ctitelka	60
cvičitelka	52
čekatelka	12
dopisovatelka	13
doručovatelka	50
držitelka	177
hlasatelka	114
hostitelka	352
chovatelka	53
jednatelka	121
krotitelka	59
léčitelka	75
majitelka	1321
nakladatelka	22
navrhovatelka	40
ničitelka	14
nositelka	178
obdivovatelka	89
objednatelka	12
obyvatelka	406
opatrovatelka	16
ošetřovatelka	879
oznamovatelka	16
pachatelka	54
pastelka	236

Produktivita

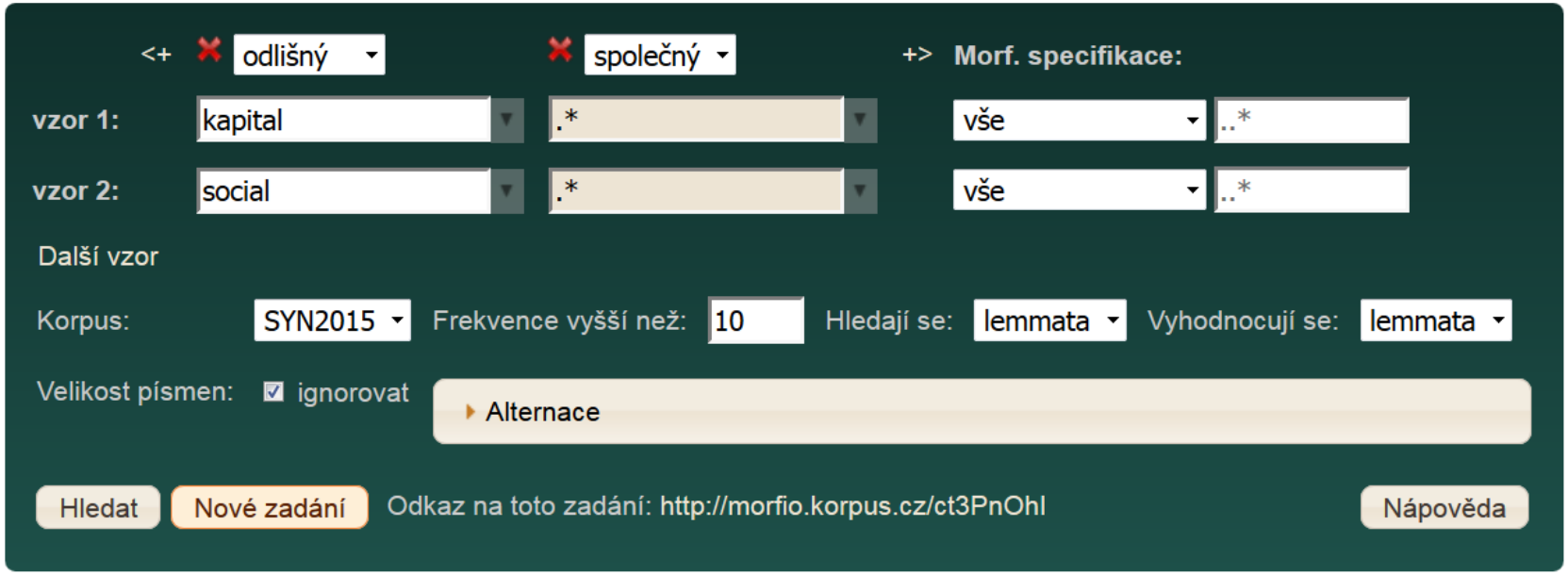


Měření produktivity

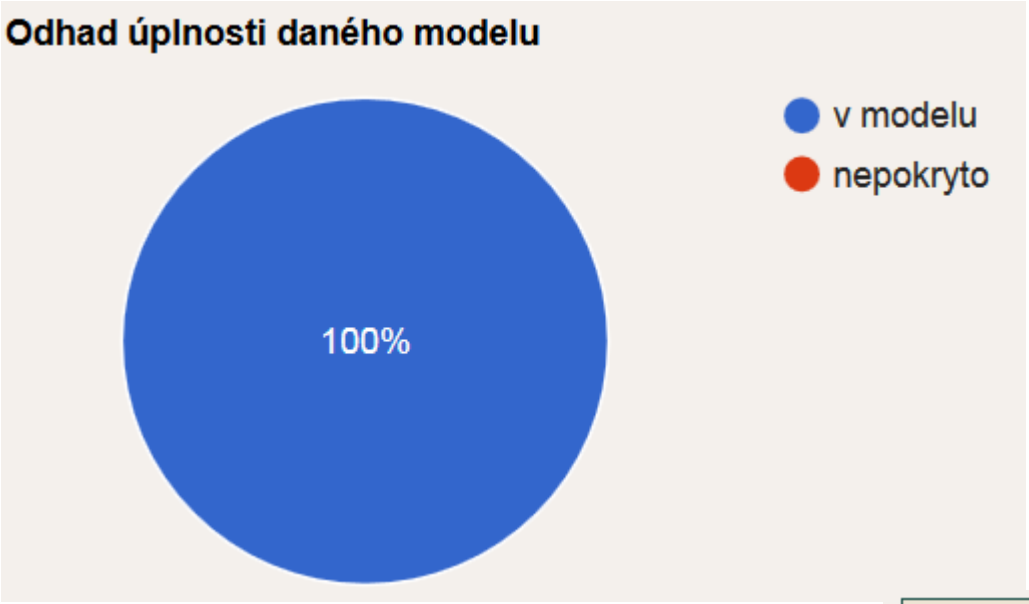
- Odhad produktivity obou vzorů a jejich vzájemné porovnání vychází z teoretických poznámek H. Baayena (viz zde). Morfologická produktivita se zde měří pomocí odhadu tendence přírůstku nových typů při přírůstku dokladů (tokenů) pro každý vzor samostatně. Ze srovnání pak vyplývá, který vzor je produktivní, protože počet jeho typů roste rychleji, s jeho formanty se pojí nové a nové báze, a který vzor je naopak neproduktivní a potenciálně uzavřený (i když třeba frekventovaný a rozsáhlý).

kapital.*/social.*

Morfio

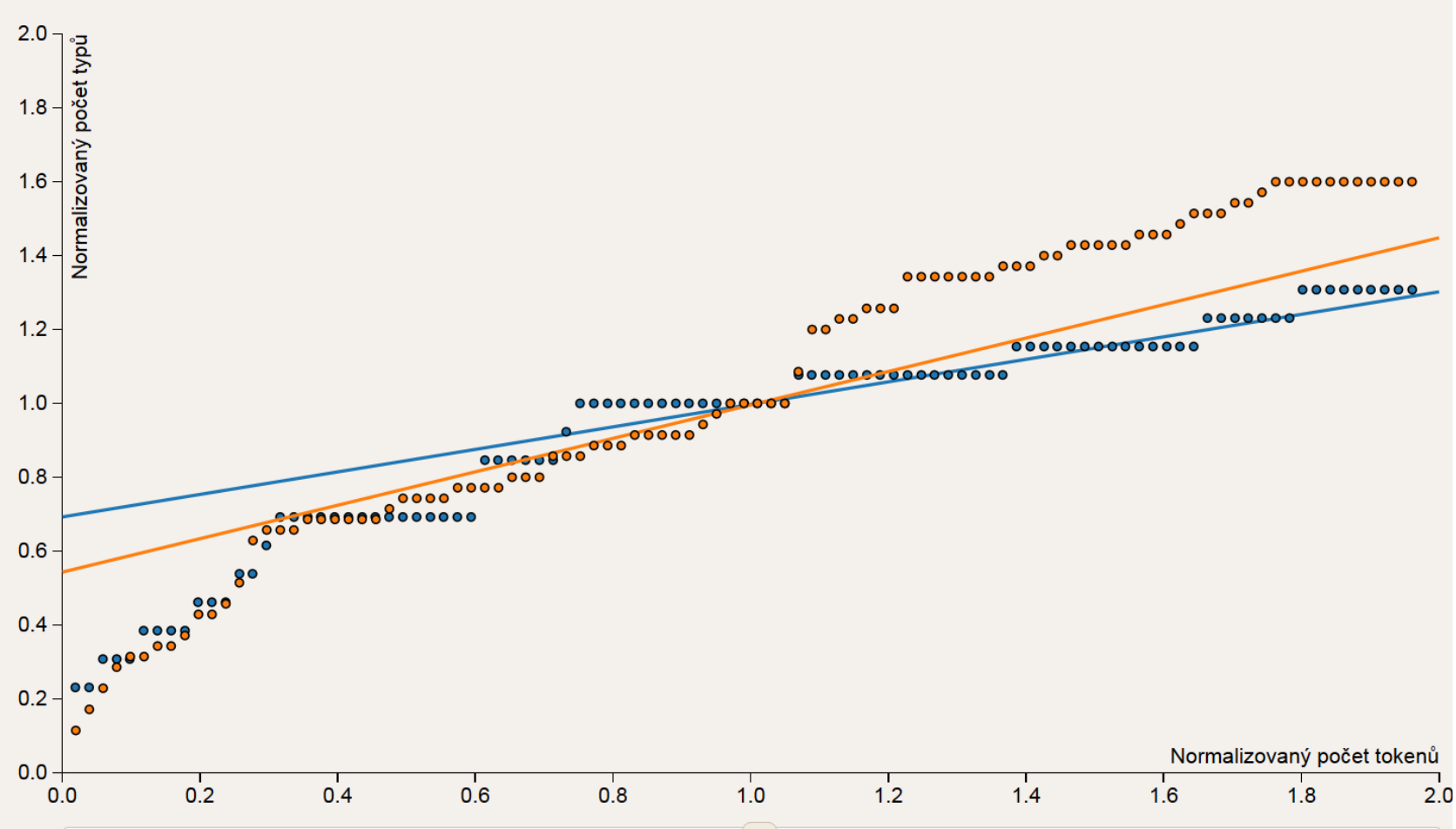


Souhrn Výpis



	vzor 1 (fq)	vzor 2 (fq)
1	kapitalismus (1011)	socialismus (1497)
2	kapitalista (203)	socialista (989)
3	kapitalistický (538)	socialistický (2446)
4	kapitalizace (68)	socializace (399)

produktivita



Vzor 1

skupiny +/-

abc/cba	fq
kapitalismus	1011
kapitalista	203
kapitalistický	538
kapitalizace	68

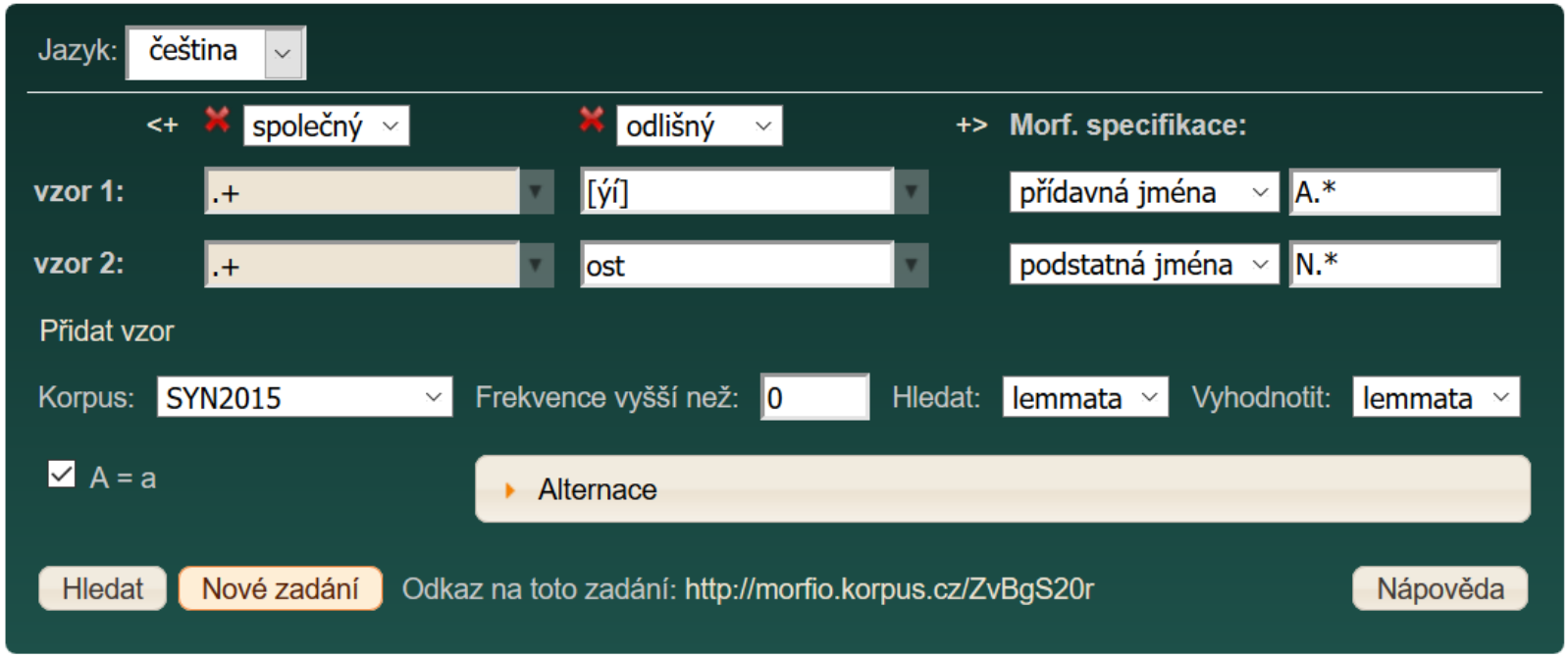
Vzor 2

skupiny +/-

abc/cba	fq
social	190
socialbakers	31
socialismus	1497
socialista	989
socialistickorealistický	11
socialisticky	27
socialistický	2446
socialistka	20
socializace	399
socializační	74
socializovaný	29
socializovat	39
socializující	13

pomocí aplikace morfio vyhledejte v korpusu syn2015 dvojice základové slovo=adjektivum, odvozené slovo= substantivum název vlastnosti

Morfio



seznam podle frekvence adjektiv

	vzor 1 (fq)	vzor 2 (fq)
1	velký (190639)	velkost (18)
2	nový (137415)	novost (160)
3	celý (133678)	celost (73)
4	jiný (124171)	jinost (3)
5	dobrý (116272)	dobrost (3)
6	český (102231)	českost (30)
7	malý (94847)	malost (133)
8	vysoký (78490)	vysokost (6)
9	starý (69518)	starost (9779)
10	vlastní (59617)	vlastnost (12735)
11	možný (52381)	možnost (43140)
12	mladý (48961)	mladost (72)
13	různý (48517)	různost (205)
14	jediný (46240)	jedinost (8)
15	stejný (41494)	stejnost (49)
16	důležitý (40918)	důležitost (1766)
17	známý (34311)	známost (1073)
18	evropský (31600)	evropskost (3)
19	plný (30642)	plnost (270)
20	špatný (30445)	špatnost (209)

přidán variantní sufix -ota

seznam podle substantiv na -ota

	vzor 1 (fq)	vzor 2 (fq)	vzor 3 (fq)
1	hodný (5344)	hodnost (841)	hodnota (29676)
2	teplý (8817)	teplost (1)	teplota (15213)
3	temný (7125)	temnost (12)	temnota (2661)
4	prázdný (11068)	prázdnost (9)	prázdnota (1618)
5	dobrý (116272)	dobrost (3)	dobrota (1322)
6	nahý (4180)	nahost (5)	nahota (461)
7	slepý (3046)	slepost (4)	slepota (352)
8	nový (137415)	novost (160)	novota (331)
9	prostý (5544)	prostost (1)	prostota (262)
10	mastný (1572)	mastnost (1)	mastnota (184)
11	hluchý (722)	hluchost (3)	hluchota (140)
12	němý (1214)	němost (3)	němota (133)
13	živý (12759)	živost (124)	života (119)
14	černý (29197)	černost (3)	černota (106)
15	cizí (11286)	cizost (47)	cizota (64)
16	milý (7399)	milost (2209)	milota (45)
17	měkký (5206)	měkkost (191)	měkkota (37)
18	teskný (137)	tesknost (2)	tesknota (36)
19	slabý (9391)	slabost (2027)	slabota (25)
20	tichý (6918)	tichost (618)	tichota (18)
21	malý (94847)	malost (133)	malota (9)
22	suchý (6382)	suchost (56)	suchota (2)
23	zlý (7155)	zlost (1270)	zlota (2)
24	mladý (48961)	mladost (72)	mladota (1)
25	mokrý (3610)	mokrost (1)	mokrota (1)

OTÁZKY

- Sledujte doklady, které se vám z nějakého důvodu zdají nepatřičné.
- Zobrazte si jejich výskyty a snažte se najít postup, jak přeformulovat dotaz tak, aby výsledek nebyl přegenerovaný.

Úkol na příště

- Pomocí aplikace *MORFIO* vyhledejte v korpusu SYN2010 kandidáty na deriváty substantiv tvořených ze sloves sufixem *–č*, přičemž chceme pouze názvy živých bytostí (hrát/hráč), takže např. dvojice jako *vařit-vařič* nás zajímá pouze v případě, že jde u vzoru dvě o označení člověka vyrábějícího pokoutně drogy a nikoli o neškodnou část běžného kuchyňského vybavení.
- Popište problémy, na které jste při práci narazili a připravte si dotazy k technickým problémům.