

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:morfologicka_analyza)
- <http://wiki.korpus.cz/doku.php/pojmy:tag>

Jak vypadá přístup

The screenshot shows the Morfio web application interface. At the top, the browser address bar displays "morfio.korpus.cz/#". Below the address bar, there are navigation links: "Nejnavštěvovanější", "Jak začít", and "Přehled zpráv". The main header features the "Morfio" logo and a red box labeled "TESTOVACÍ PROVOZ".

The search interface includes several filters and options:

- Left side: "<+" button, a red "X" icon, a dropdown menu set to "společný", another red "X" icon, and a dropdown menu set to "odlišný".
- Right side: "+>" button and "Morf. specifikace:" label.
- Search examples: "vzor 1:" and "vzor 2:" each followed by a ".+" dropdown, an empty dropdown, a "vše" dropdown, and a "..*" text input.
- Additional options: "Další vzor" label, "Korpus:" dropdown set to "SYN2010", "Frekvence vyšší než:" input set to "10", "Hledají se:" dropdown set to "lemmata", and "Vyhodnocují se:" dropdown set to "lemmata".
- Character size: "Velikost písmen:" with a checked "ignorovat" checkbox and a large "Alternace" button.
- Bottom navigation: "Hledat" button, "Nové zadání" button, a checkbox for "zachovat záznam procesu", and "Nápověda" button.

Volba korpusu

- SYN2010
- SYN2005

Zadání dotazu

- regulární výrazy
- morfologické značky (viz výše)
- další možnosti zobecnění (rozlišení samohlásek, souhlásek, ...)
- 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

← morfio.korpus.cz/#

Nejnavštěvovanější Jak začít Přehled zpráv

Morfio **TESTOVACÍ PROVOZ**

<+ společný odlišný +> Morf. specifikace:

vzor 1:

vzor 2:

Další vzor

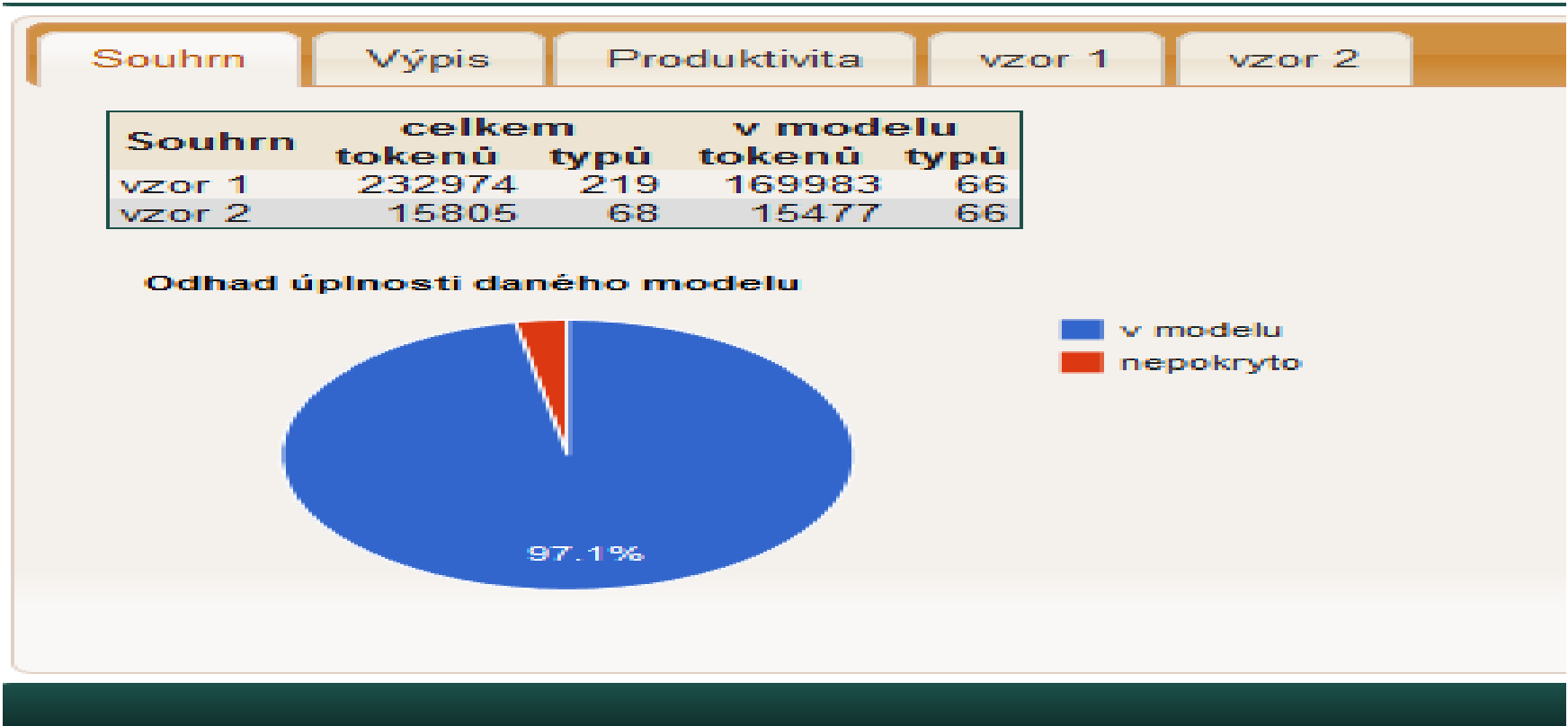
Korpus: Frekvence vyšší než: Hledají se: Vyhodnocují se:

Velikost písmen: ignorovat

zachovat záznam procesu

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 slovtvorného 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="(?)badatel" & tag="NNM.*"]" z korpusu syn2010

Fuks, Ladislav: Vévodkyně a kuchařka

Strachey, Lytton: Alžběta a Essex

Y: Austrálie

Gieselmann, David: Pan Kolpert

Mácha, Karel: Integrovaná antropolo...

Vacková, Jarmila: Odpovědi obrazů - Mi...

Gurevič, Aron Jakovle...: Nebe, peklo, svět - ...

Gurevič, Aron Jakovle...: Nebe, peklo, svět - ...

Brunschvicg, Léon: Evropský duch

Bauman, Zygmunt: Úvahy o postmoderní době

Potůček, Martin (ed.): Manuál prognostickýc...

Keller, Jan: Teorie modernizace

Y: Ikarie, č. 4/2005

Studnička, Miloslav: Masožravé rostliny

Benda, Ivo A. (et al.): VI. rozhovory s pouč...

Y: Technik, č. 7/2005

Koukolík, František -...: Šimpanz a vesmír

Štoll, Ivan: Podivuhodné přírodní úkazy

Wolf, Josef: Kdo je kdo není v hn...

Y: S tebou mě baví život, č. 2/2008

Kalina, Pavel - Koťát...: Praha 1310-1437

Y: Právo, 3. 9. 2005

Y: Lidové noviny, 17. 7. 2006

Y: Mladá fronta DNES, 25. 4. 2007

Y: Sedmička, č. 18/2009

společníci - kapitán navíc jistý **badatel** , cestovatel a znalec Afriky
 úzkostném napětí a některé moderní **badatele** vedla k domněnce , že
 Průkopníci a badatelé Mnoho evropských **badatelů** , včetně Edwarda Eyra a
 takovou naprosto banální odpověď : **badatel** v oblasti chaosu - tedy
 snad zdát tato formulace konkrétnímu **badateli** v některé přírodovědní disciplíně příliš
 Malerei (1925) německý **badatel** M. J. Friedländer s tím
 s tvou služebnicí ? " **Badatelé** spatřují v rituálu " snižování
 Evropě 13 . století odhalují **badatelé** na několik stovek , vedlo
 : " Cítím se být **badatelem** . Žízním po úplném poznání
 městské kultury ; pro mnoho **badatelů** se stal zevloun hlavním symbolem
 souvislosti , které nejsou zřejmé **badatelům** v individuálních výzkumných oblastech .
 sponzorů ztrácejí univerzitní učitelé a **badatelé** status nezávislých odborníků . Část
 Cristea , občan Ukrajiny a **badatel** . Měl jednu bodnou ránu
 darlingtonii a pavoukem zmiňují různí **badatelé** , včetně historických zpráv zmíněné
 Uvědomte si proto , pozemští **badatelé** , že to , co
 . let dvacátého století se **badatelům** nepodařilo dostat se při snižování
 říkají stochastickými - vlivy . **Badatelé** zkoumající dlouověkost drosofil se samozřejmě
 1597) , holandský polární **badatel** , když v roce 1596
 se vyvíjela od původního kroužku **Badatelů** Bible v sedmdesátých letech 19
 F. X. Šalda . Literární **badatel** Radko Pytlík však shrnul spisovatelovo
 jižní pilíř unikál pozornosti novodobých **badatelů** . Je to patrně dáno
 z pozdní doby bronzové našel **badatel** spolupracující s mladoboleslavským muzeem v
 je nezískal vůbec . Cambridgeští **badatelé** řadu let zkoumali třináct skupin
 poškozené spisy neměly být přístupné **badatelům** . " Můžeme ho ukázat
 , spoustu zajímavých informací však **badatelé** objev 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 stejnouází, lišící se pouze formanty).

tel (vzor 1)

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

skryt 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
čítatel	32
datel	135
dělitel	13
dobyvatel	441
dodavatel	4822
dohlížitel	30
dohlížitel	61
dopisovatel	188
doručitel	40
doručovatel	136
dovolatel	14
dražitel	73
držitel	1729
expřitel	31
exředitel	28
gerstel	11

telka (vzor2)

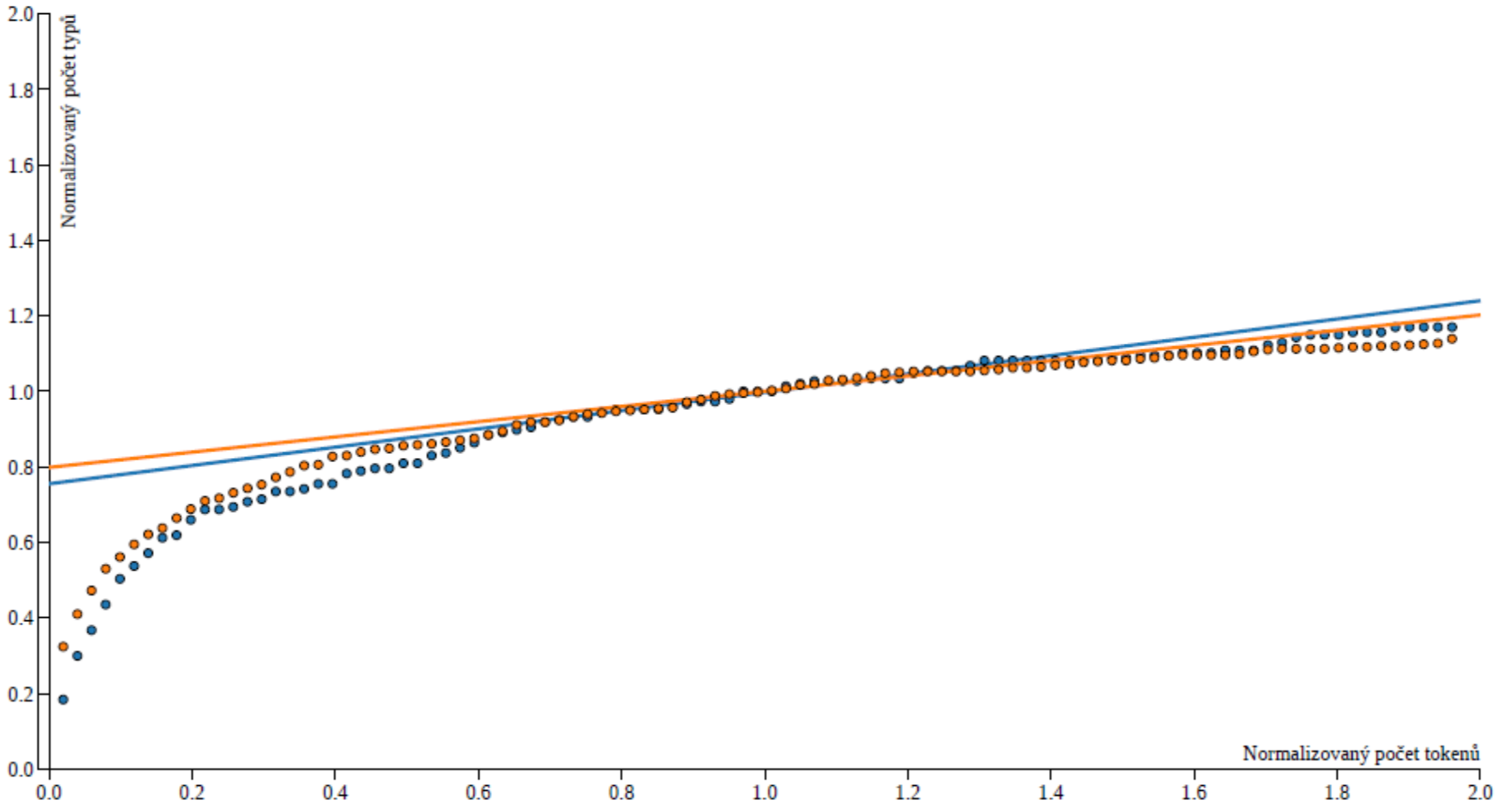
- Žlutě jsou ty členy, k nimž byl nalezen vzor 1

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).

Úkol na 28. 10. 2013

- 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-vaříč* 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**.