PLIN033

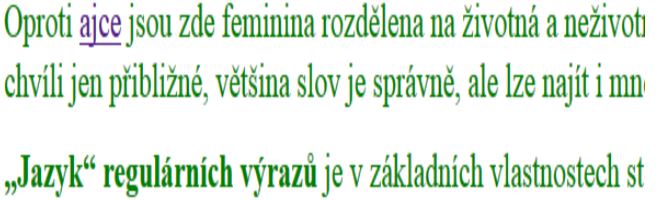
Deriv – nástroj pro testování derivačních vztahů ve strojovém slovníku a v korpusech

Co dnes chceme?

- představit některé funkce *Deriv*
- ukázat na konkrétním příkladu postup práce při extrakci podkladů pro lingvistickou analýzu slovotvorných formací automaticky získaných prostřednictvím *Derivu*
- zadat úkoly na příště

Co je to *Deriv?*

- webové rozhraní
- schopnost pracovat s morfologickým slovníkem analyzátoru *(m)ajka*
- propojení s tištěnými slovníky
- propojení s korpusy


Oproti ajce jsou zde feminina rozdělena na životná a neživot
chvíli jen přibližné, většina slov je správně, ale lze najít i mn
„Jazyk“ regulárních výrazů je v základních vlastnostech st

ajka – tagset

- http://nlp.fi.muni.cz/projekty/ajka/tags.pdf

ajka tagset

Radek Sedláček

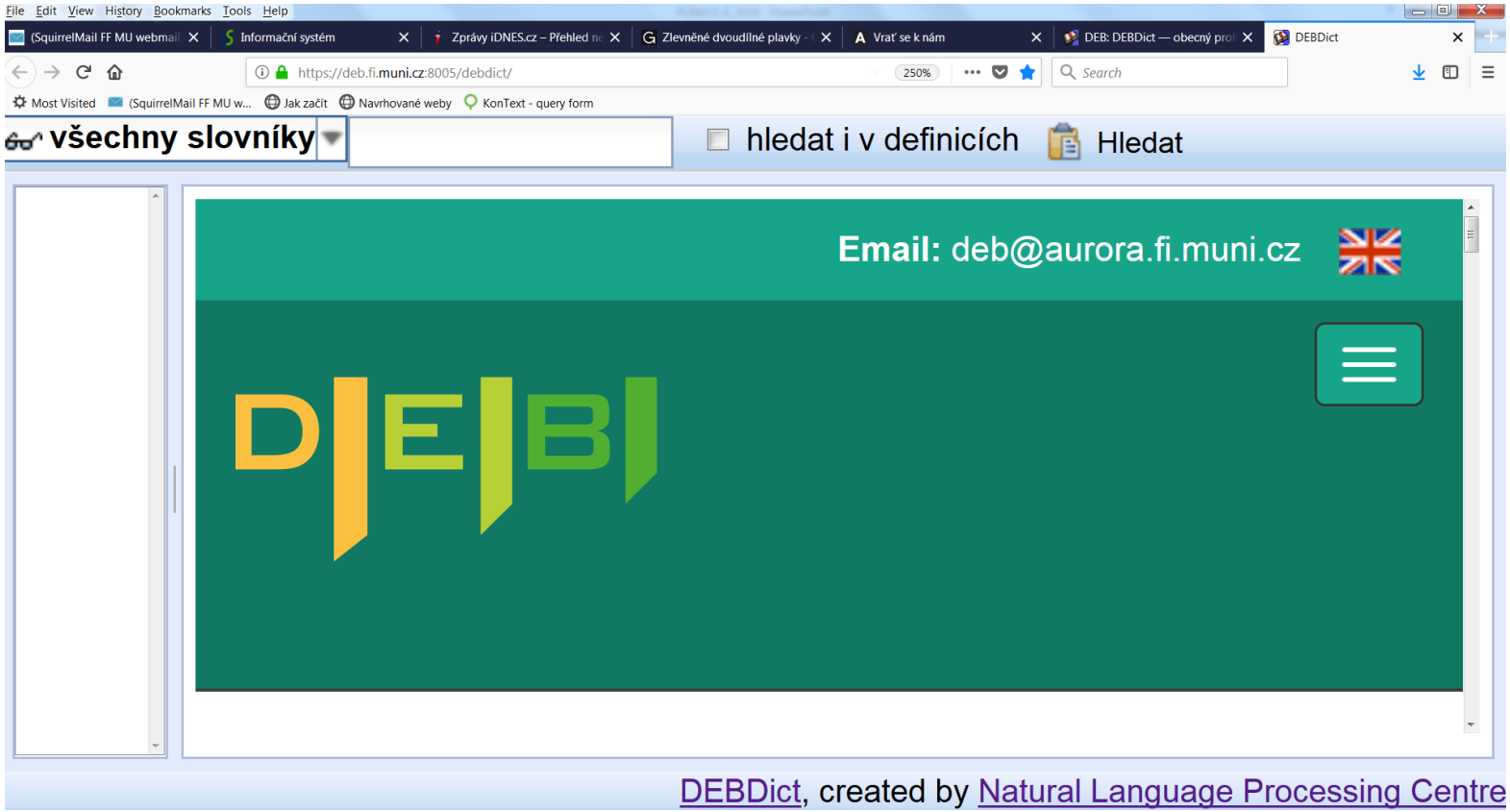
March 1, 2006

1 k1 - Substantivum

x	Speciální vzor
P	půl

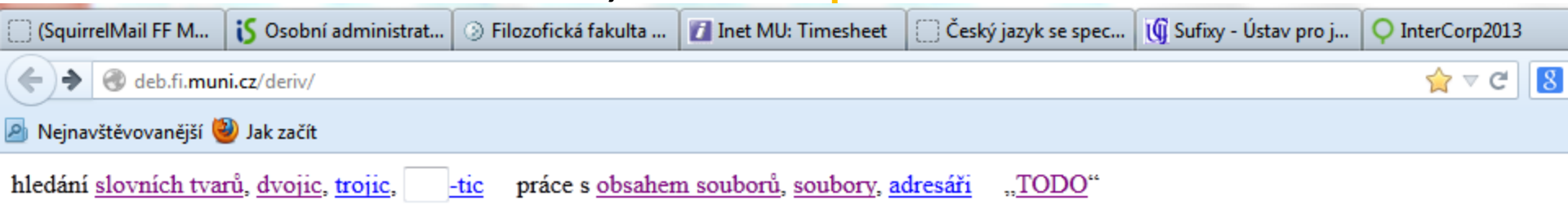
g	Rod
M	Rod mužský životný
I	Rod mužský neživotný
N	Rod střední
F	Rod ženský
R	Rodina (příjmení)

debdict


File Edit View History Bookmarks Tools Help
DEBDict
https://deb.fi.muni.cz:8005/debdict/
250%
Search
Most Visited
Jak začít
Navrhované weby
KonText - query form
všechny slovníky
hledat i v definicích
Hledat
Email: deb@aurora.fi.muni.cz
DEBDict, created by Natural Language Processing Centre

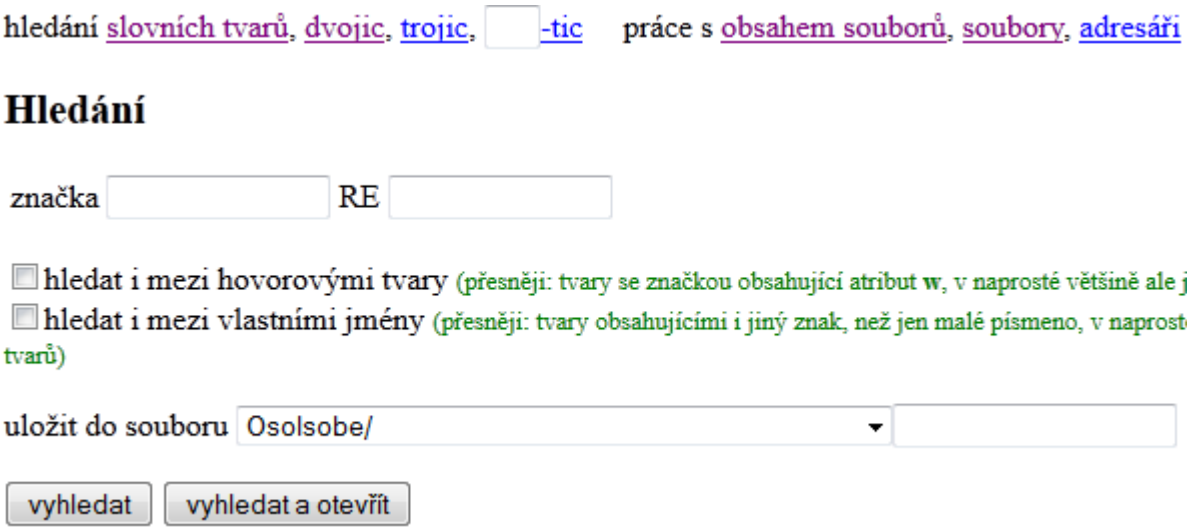
deb.fi.muni.cz/deriv

- HESLO: smerk@mail.muni.cz.
- Jméno: PLIN033, heslo: plin



Hledání slovních tvarů

- Lze hledat dle značky (brněnský systém - http://nlp.fi.muni.cz/projekty/ajka/tags.pdf)



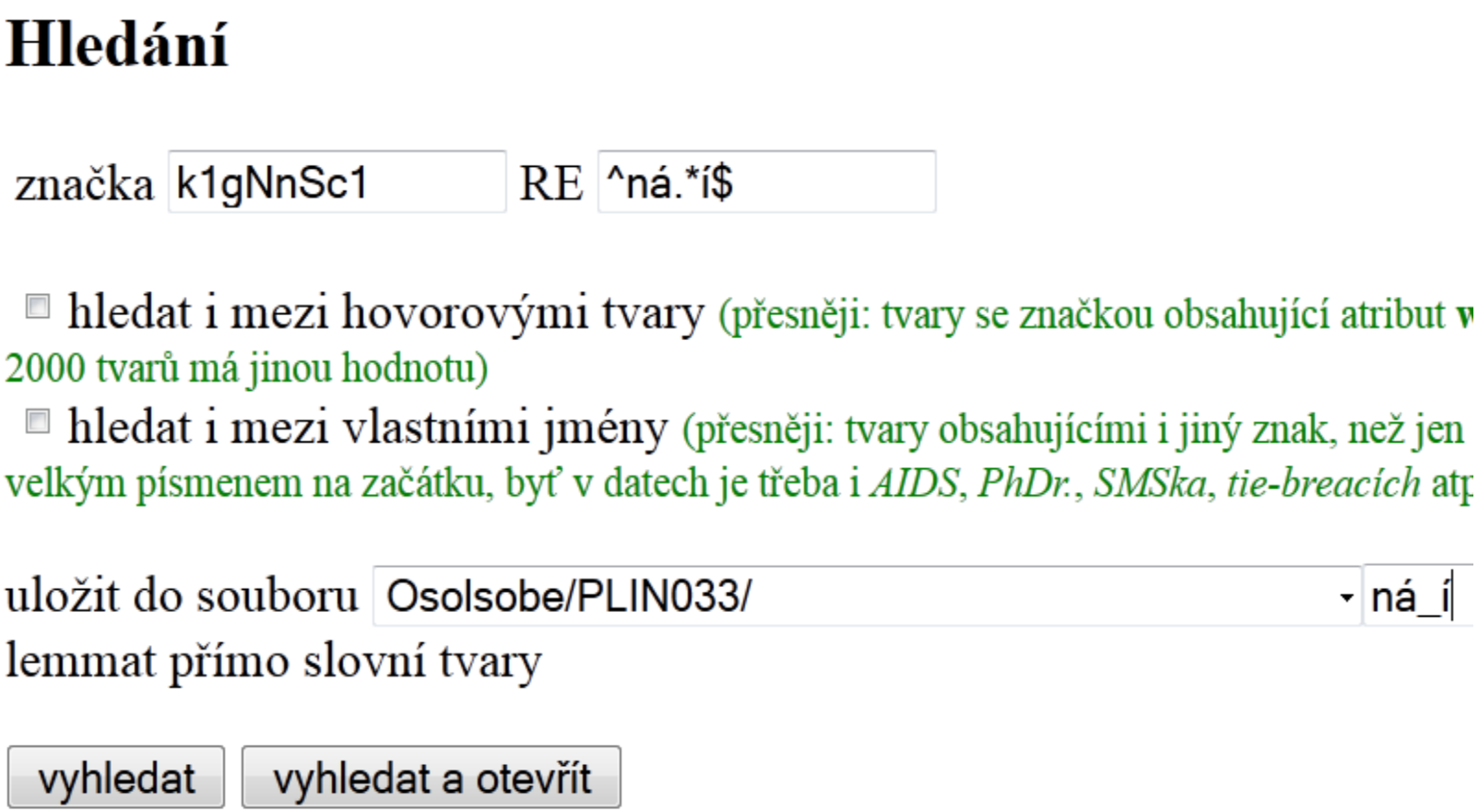
Co chceme?

- Substantiva typu *náměstí*
- Jaké další známe?
- Jak je můžeme popsat?
- Jaké mají formální vlastnosti?

ná- -í

- Substantiva
- Neutra
- Lemma začíná na *ná* a končí na *í*

Vyplnění formuláře



Vyhledat a otevřít

- **Seznam slov, která odpovídají zadání a která jsou uvedena ve slovníku morfologického analyzátoru *ajka***

Práce s obsahem souboru PLIN033/ná_í

Soubor **PLIN033/ná_í** byl úspěšně uložen (0 s).

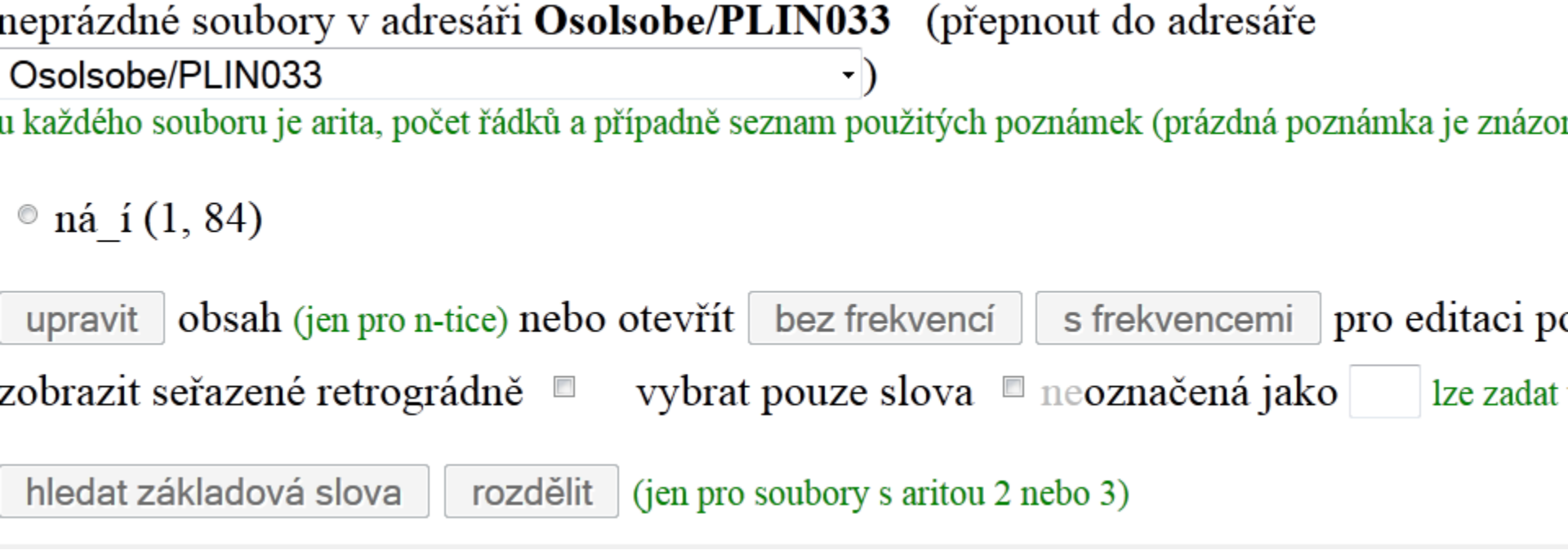
Poznámkou je jednotlivý znak, jeden řádek může mít více poznámek.
Jako poznámku nelze použít dvojtečku, čárku a mezeru (budou-li zadány, budou ignorovány).

1 náboženství
2 nábožnůstkářství
3 nábřeží
4 nábytkářství
5 náčelnictví
6 náčiní
7 nádbí
8 nádenictví
9 nádeničení
10 nádní

Co lze dále zjistit?

- Otevřít soubor
- Upravovat jeho obsah
- Prohlížet frekvence
- Nahlížet do slovníků
- Nahlížet do korpusů

Otevření s frekvencemi



slovo (frekvence)

Práce s obsahem souboru PLIN033/ná_í

Poznámkou je jednotlivý znak, jeden řádek může mít více poznámek.
Jako poznámku nelze použít dvojtečku, čárku a mezeru (budou-li zadány, budou ignorovány).

První číslo v závorkách je počet výsledků korpusového dotazu `[word="slovo"]` v korpusu SYN2000, druhé č
`[lemma="slovo"]` (případně seznam takových výsledků pro různé slovní druhy, viz např. slovo dolomit) v tor
počet výsledků dotazu `[word="slovo"]` v korpusu CzTenTen a čtvrté číslo je počet výsledků dotazu `[lemma=`
Zobrazená čísla jsou předpočítaná, takže se mohou lišit od počtů udávaných korpusovým managerem (zpravi
SYN2000 proto, že nemám k dispozici zdrojovou podobu verze, kterou odkazuji, u CzTenTen proto, že korpu
měnit. Druhé a čtvrté číslo nemusejí mít dobrý smysl v případech, že v souboru nejsou lemmata, ale tvary, ov
neuchovává informace, jakým způsobem soubor vznikl, takže nelze spolehlivě poznat, jestli jde o lemmata, n
zároveň fungují jako odkazy na konkordanční seznamy příslušných výskytů. Je vhodné mít na paměti, že zatí
odpovídá přímo korpusovým datům, druhé a čtvrté pracuje s výsledky automatické lemmatizace, která ovšem
chyby.

1 náboženství (2533, 3070, 154411, 186107)
2 nábožnůstkářství (0, 0, 0, 0)
3 nábřeží (988, 1085, 25625, 28834)

Kliknutím na slovo získáme informace z tištěných slovníků uložených v databázi debdict

ssjc Slovník spisovného jazyka českého

nábřeží

-í s. *upravená a podezděná městská cesta podle řeky, jezera:* vltavské n.; procházet se p
kapitána Jaroše; --> expr. zdrob. **nábřežíčko,** -a s. (6. mn. -ách)

---

ssc Slovník spisovné češtiny

nábřeží

-í s *stavebně upravený prostor podél řeky ap.* vltavské nábřeží, chodit po nábřeží,
nábřežní příd.

Kliknutím na slovo získáme informace o výskytu v korpusech

smilide	na vlhký obzor . Po nábřeží hopkaly veverky , které si
smilide	trocho anarchisticky vyhlížející mladík na nábřeží v lehké teplákové soupravě a
smilide	firmy Kodak . Jinak bylo nábřeží prázdné a mladíkovu opuštěnost umocňoval
smilide	. Nohy ho donesly na nábřeží a odtud do hospody na
smilide	, kam se dostal od nábřeží . Stále ho nikdo nesledoval
smilide	na chvíli se z okraje nábřeží zahleděl na doutnající tok řeky
smilide	. Šli spolu kousek po nábřeží , téměř až ke kříži
pi210	aby se postavil dům na nábřeží Vltavy , Hlahol . .
havel	2000 v Praze na původním nábřeží Palackého , později Rašínově ,
havel	Do nově zařízeného bytu na nábřeží č . 2000 přinesli z
havel	kolegiem . Doma , na nábřeží , mu ale chybějí .
havel	a potom zařízení bytu na nábřeží . Líbí se mi ,
ing	Dame zní zvon a po nábřeží promenují podezřelé krásky . Porci
brizy	ve čtvrtém poschodí , na nábřeží Marxe a Engelse . Mohl
svedeny	jakém smyslu ? Šli po nábřeží . Zastavil se a oběma
s_mloky	, kteří se trousí po nábřeží a po mostě , nemají
s_mloky	mostě cinkala tramvaj , po nábřeží putovaly chůvy s kočárky a
krakatit	Prokop si razí cestu po nábřeží . Mrazí ho a čelo
honzlova	jsem se dál směrem k nábřeží . Sfingy , co hlídají
honzlova	na ní rozhojnily , ale nábřeží bylo stejně mrtvé jako ráno

strana 1 ze 52 Přejít další | poslední

Jak dále ?

- Méně obvyklá a méně frekventovaná slova

16		náhlaví (0, 0, 1, 2)
17		náhlení (0, 0, 2, 2)
18		náhradbí (0, 0, 3, 4)
19		náhradí (0, 0, 151, 208)

náhlaví

ssjc Slovník spisovného jazyka českého

náhlaví

-í s. zeměd. *část předku pluhu spočívající mezi plužňaty; náhlavník*

---

ssc Slovník spisovné češtiny

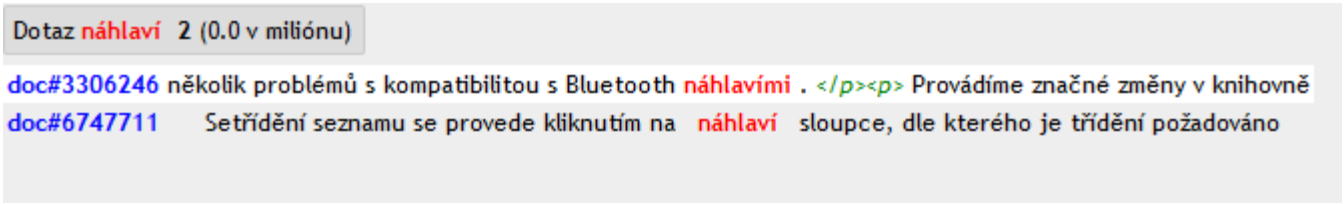
---

psjc Příruční slovník jazyka českého

náhlaví, -í n. *hosp. část předku pluhu spočívající na drábcích mezi plužňaty, náhlavník*

Výskyt v korpusu

- Odpovídá význam výskytů nalezených v korpusu významům, které uvádějí tištěné slovníky ?



Pohled do zdrojů

- Kliknutím na url lze získat u velkých korpusů z webu přístup ke zdrojovým textům

Celý text

- Co je to náhlaví?

Zde jsou nějaké věci, kterými jsme se mezitím zabývali:

- Kompletně jsme přepsali knihovnu pro práci s audiem a věnovali značné úsilí přizpůsobování stávajícího kódu pro práci se zvukem do podoby lépe vyhovující pro Skype.
  Také jsme opravili několik problémů s kompatibilitou s Bluetooth náhlavími.
- Provádíme značné změny v knihovně pro práci s videem. Pracujeme na tom, aby byla stabilnější a podporovala více webkamer a X video módů.
- Chystáme se adoptovat některá vylepšení uživatelského rozhraní ze Skype 4.0 pro Windows, například záložku s konverzacemi. Také rozmýšlíme nad přidáním jednookneního režimu.
  Samozřejmě, toto je stále ve vývoji, ale náš cíl je vytvořit jednodušší uživatelské rozhraní vyžadující méně kliknutí myši pro provedení nejběžnějších operací.

Lingvistická analýza automaticky vyhledaných dat

- Které jednotky nepatří do seznamu substantiv, neuter tvořených cirkumfixem *ná- -í*?
- Projděme seznam (84 jednotek).
- Využijme funkce nástroje *Deriv* označovat nalezené jednotky (okénko mezi pořadovým číslem a slovem).

Ruční práce

- Co je to *nádbí* a *nádní*?

1	X	náboženství (2533, 3070, 154411, 186107)
2	X	nábožnůstkářství (0, 0, 0, 0)
3		nábřeží (988, 1085, 25625, 28834)
4	X	nábytkářství (19, 20, 551, 600)
5	X	náčelnictví (4, 5, 124, 151)
6		náčiní (323, 453, 14873, 17969)
7		nádbí (0, 0, 7, 7)
8	X	nádenictví (0, 0, 12, 18)
9	Y	nádeničení (0, 0, 12, 6)
0		nádní (0, 0, 30, 42)
1		nádobí (1352, 1492, 69703, 76543)

nádbí

- Odpovídá výklad ve slovníku významu korpusových vyhledávek?

psjc Příruční slovník jazyka českého

nádbí, -í n. *dial.* naděje nádba Což honem nevěstu. Byla v nádbí, Bortková „pantlova
vyhlédnuta. A. Mrš.

Dotaz nádbí 7 (0.0 v miliónu)

doc#29780	(Oh My Fucking God - jak se jednorázové nádbí tohoto typu vůbec do takové kuchyně dostane
doc#423144	mraznička, mikrovlnka, rychlovarná konvice, nádbí). V apartmá je připojení na internet, televize
doc#1497507	vysoké úrovni. Ať už je to keramika, dřevěné nádbí , košíky, hračky nebo dokonce i bylinky,
doc#1517984	když si spočítáte, jak dlouho vám takové nádbí vydrží, tak vám jistě dojde, co všechno
doc#1699136	dobře čísti nedovedeme (w nadby, snad = v nádbí). Pokus vyložiti je by našel čtenář ve
doc#2060149	odmastí i organismus, protože ho tak úplně z nádbí nesmyješ, takže s dalším jídlem konzumuješ
doc#7072646	přepracovat dané receptury a mít dostatečně velké nádbí . Obyvatelé a návštěvníci Domažlic

nádní

nádní I

příd. loď. n. prostor *jsoucí přímo nade dnem lodi*

nádní II

v. návní

SSC Slovník spisovné češtiny

PSJC Příruční slovník jazyka českého

nádní adj. **1.** *loď.* nádní prostor *který je přímo nade dnem lodi.*

nádní adj. **2.** *v.* návní

nádní, -í n. *spodek lodi.* Tesal nádní loďky nové. Hol. Loď kde nádním smýká mžiká. F. S. Proch.

nádní

- Odpovídá značkování?

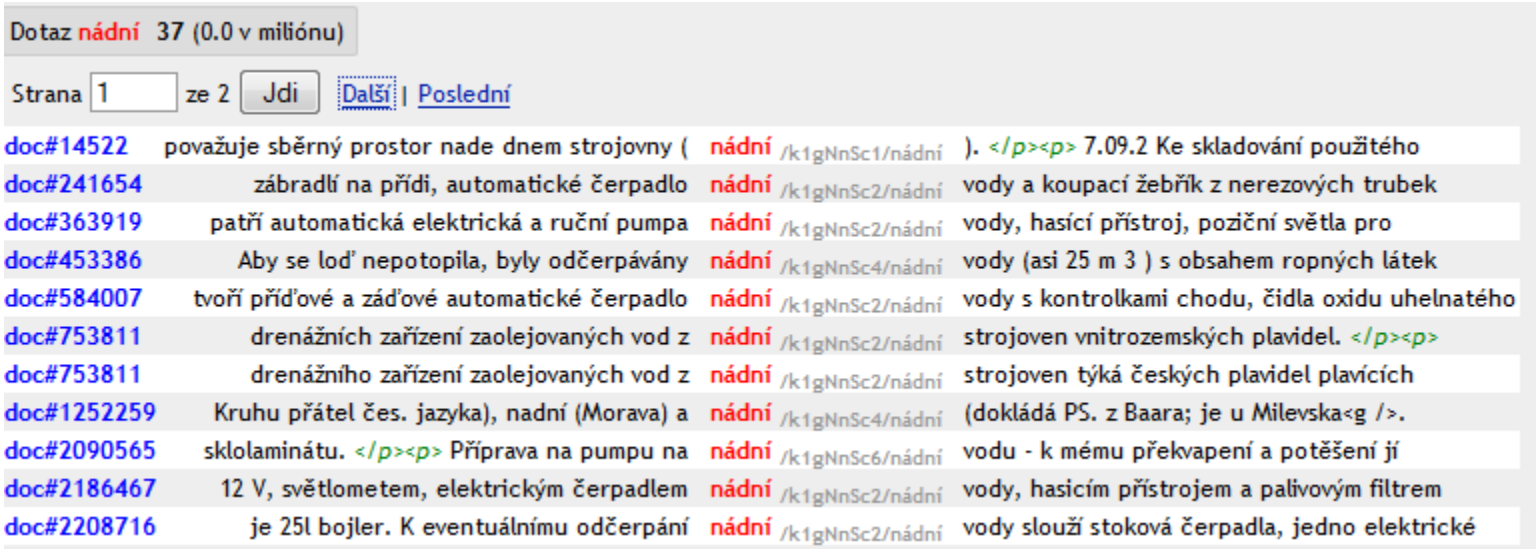
Dotaz nádní 37 (0.0 v miliónu)

Strana 1 ze 2 Jdi Další | Poslední

doc#14522	považuje sběrný prostor nade dnem strojovny (	nádní	). 7.09.2 Ke skladování použitého
doc#241654	zábradlí na přídi, automatické čerpadlo	nádní	vody a koupací žebřík z nerezových trubek
doc#363919	patří automatická elektrická a ruční pumpa	nádní	vody, hasící přístroj, poziční světla pro
doc#453386	Aby se loď nepotopila, byly odčerpávány	nádní	vody (asi 25 m 3) s obsahem ropných látek
doc#584007	tvoří příďové a záďové automatické čerpadlo	nádní	vody s kontrolkami chodu, čidla oxidu uhelnatého
doc#753811	drenážních zařízení zaolejovaných vod z	nádní	strojoven vnitrozemských plavidel.
doc#753811	drenážního zařízení zaolejovaných vod z	nádní	strojoven týká českých plavidel plavících
doc#1252259	Kruhu přátel čes. jazyka), nadní (Morava) a	nádní	(dokládá PS. z Baara; je u Milevska<g />.
doc#2090565	sklolaminátu. Příprava na pumpu na	nádní	vodu - k mému překvapení a potěšení jí
doc#2186467	12 V, světlometem, elektrickým čerpadlem	nádní	vody, hasicím přístrojem a palivovým filtrem
doc#2208716	je 25l bojler. K eventuálnímu odčerpání	nádní	vody slouží stoková čerpadla, jedno elektrické
doc#2264427	vybaveny měřičem aktuálního stavu. Čerpadla	nádní	vody jsou celkem tři - ruční, elektrické

Chyby ve značkách

- I adjektivní výskyty jsou označkovány jako substantiva



Dotaz nádní 37 (0.0 v miliónu)

Strana 1 ze 2 Jdi Další | Poslední

doc#14522	považuje sběrný prostor nade dnem strojovny (	nádní	/k1gNnSc1/nádní	). 7.09.2 Ke skladování použitého
doc#241654	zábradlí na přídi, automatické čerpadlo	nádní	/k1gNnSc2/nádní	vody a koupací žebřík z nerezových trubek
doc#363919	patří automatická elektrická a ruční pumpa	nádní	/k1gNnSc2/nádní	vody, hasící přístroj, poziční světla pro
doc#453386	Aby se loď nepotopila, byly odčerpávány	nádní	/k1gNnSc4/nádní	vody (asi 25 m 3) s obsahem ropných látek
doc#584007	tvoří příďové a záďové automatické čerpadlo	nádní	/k1gNnSc2/nádní	vody s kontrolkami chodu, čidla oxidu uhelnatého
doc#753811	drenážních zařízení zaolejovaných vod z	nádní	/k1gNnSc2/nádní	strojoven vnitrozemských plavidel.
doc#753811	drenážního zařízení zaolejovaných vod z	nádní	/k1gNnSc2/nádní	strojoven týká českých plavidel plavících
doc#1252259	Kruhu přátel čes. jazyka), nadní (Morava) a	nádní	/k1gNnSc4/nádní	(dokládá PS. z Baara; je u Milevska<g />.
doc#2090565	sklolaminátu. Příprava na pumpu na	nádní	/k1gNnSc6/nádní	vodu - k mému překvapení a potěšení jí
doc#2186467	12 V, světlometem, elektrickým čerpadlem	nádní	/k1gNnSc2/nádní	vody, hasicím přístrojem a palivovým filtrem
doc#2208716	je 25l bojler. K eventuálnímu odčerpání	nádní	/k1gNnSc2/nádní	vody slouží stoková čerpadla, jedno elektrické

Zvolme značky a vyznačme případy tzv. přegenerování

- XY?

1	X	náboženství (2533, 3070, 154411, 186107)
2	X	nábožnůstkářství (0, 0, 0, 0)
3		nábřeží (988, 1085, 25625, 28834)
4	X	nábytkářství (19, 20, 551, 600)
5	X	náčelnictví (4, 5, 124, 151)
6		náčiní (323, 453, 14873, 17969)
7	?	nádbí (0, 0, 7, 7)
8	X	nádenictví (0, 0, 12, 18)
9	Y	nádeničení (0, 0, 12, 6)
10		nádní (0, 0, 30, 42)
11		nádobí (1352, 1492, 69703, 76543)
12	?	nádrabí (0, 0, 0, 0)
13		nádraží (4591, 5283, 208635, 235714)
14	X	nádvornictví (0, 0, 0, 0)
15		nádvoří (1061, 1168, 50456, 55587)
16		náhlaví (0, 0, 1, 2)
17	Y	náhlení (0, 0, 2, 2)

Přegenerování a podgenerování jako obecný problém automatické analýzy

- Podmínka v zadání je nutná, nikoli dostačující.
- Důsledek – přegenerování – výsledek obsahuje data, která jsme vyhledat nezamýšleli.
- Důsledek – podgenerování – výsledek neobsahuje data, která jsme vyhledat zamýšleli.

Přegenerované výsledky

- Substantiva (kolektiva) na [sc]tví – X.
- Dějová jména od sloves začínajících na *ná-* – Y.
- Další (kompozitum *názvosloví*) – Z.
- Podezřelé (chyby ve značkování) – !
- Podezřelé (překlepy) – ?

Uložení do nového souboru

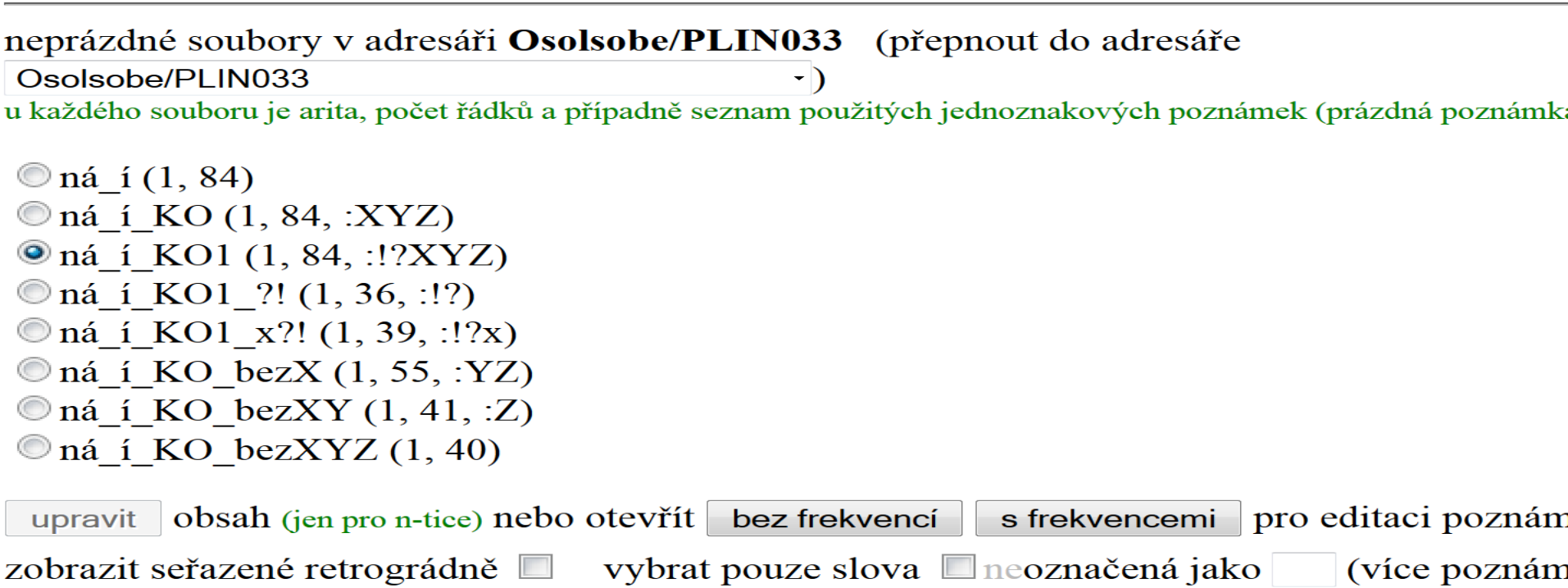
- Název nového souboru se liší od starého

76	X	nástupnictví (174, 194, 3171, 3381)
77		návěstí (41, 43, 1198, 454)
78		návěští (22, 32, 1188, 1555)
79		návětří (1, 1, 167, 185)
80		návidlí (0, 0, 0, 1)
81	X	návladnictví (0, 0, 28, 30)
82	X	návrhářství (40, 45, 1911, 2062)
83		návrší (220, 253, 9651, 5533)
84	Z	názvosloví (156, 177, 8957, 10216)

uložit do souboru Osolsobe/PLIN033/ ná_í_KO1

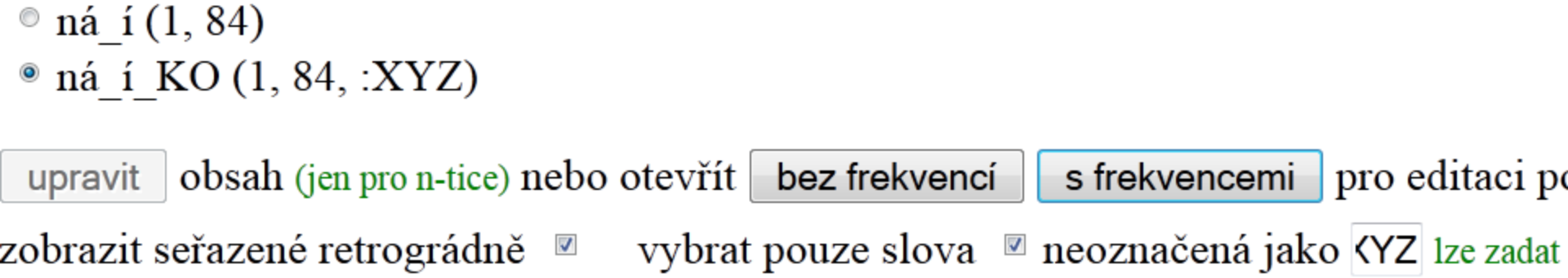
Označkovaný soubor

- Funkce vybrat pouze slova označená jako



Vybrat slova bez poznámek

- Je možné ručně upravit – odstranit přegenerovaná data.



Postup odstranění/výběru ručně označkovaných jednotek

- Chceme-li zvolit více poznámek (ručně zvolených značek), pak je třeba zadat volbu pod menu *„více poznámek spojit spojkou“.*

Soubor bez ručně označených dat

- neodstranili jsme podezřelé jednotky

2		náčiní (323, 453, 14873, 17969)
3	?	nádbí (0, 0, 7, 7)
4	!	nádní (0, 0, 30, 42)
5		nádobí (1352, 1492, 69703, 76543)
6		nádrabí (0, 0, 0, 0)
7		nádraží (4591, 5283, 208635, 235714)

Soubor uložíme

- Nový název souboru je např. ná_í_KO1_?!

35 návěstí (41, 43, 1198, 454)
36 návěští (22, 32, 1188, 1555)
37 návětří (1, 1, 167, 185)
38 návidlí (0, 0, 0, 1)
39 návrší (220, 253, 9651, 5533)

uložit do souboru Osolsobe/PLIN033/ ná_í_KO1_?!

Další zpracování

- můžeme např. odstranit „podezřelé“ jednotky stejně, jak bylo uvedeno výše
- můžeme práci odložit
- můžeme dále kontrolovat, zda jsme někde neudělali chybu
- můžeme se věnovat lingvistickému popisu, např. vytvořit následující slovníkové heslo

slovníkové heslo

- návrh

ná- -í

Stavba: *ná- -(í)*

Cirkumfixem *ná- -í* se tvoří okrajově **substantiva** od **I. substantiv.** Označují názvy **(1a) míst**, a to jednak z předložkových pádů jmen užitých v odpovědi na otázku ***kde?*** (v úzu běžná a frekventovaná *náměstí, nádraží* i okazionálně tvořená *námoří*), jednak **(1b)** názvy **míst s působením nějakého povětrnostního vlivu** (*náledí, návětří*), jednak názvy **(1c) dle polohy** (*náručí,* archaický termín *nákončí*). Synchronně většinou neprůhledné jsou názvy **(2) souborů prostředků** (*nádobí, náčiní, nářadí*). Význam **(3) uplatnění motivujícího substantiva** má *násilí*. V době obrozenecké (Dobrovský) bylo přejato utvořené desubstantivum označující **(4)** dialekt (*nářečí*). Právnický termín je **(5)** *návětí* = premisa.
Od **II. slovesa** je utvořeno substantivum s významem **(6) prostředku** (*návěstí, návěští*).

nebo se na data podívat z hlediska frekvenční analýzy

- vybrat slova, která uvadějí slovníky, ale v korpusech se nevyskytují (nebo jde o překlepy)
- vybrat slova, která mají velmi nízké frekvence (hapaxy)

Vyhledávání dvojic

- Sloveso – maskulinum činitelské jméno na *č (topit/topič)*
- **t$/k5.*mF>č/k1gMnSc1**
- **at$/k5.*mF>áč/k1gMnSc1**
- **ít$/k5.*mF>eč/k1gMnSc1**
- **[eí]t$/k5.*mF>ič/k1gMnSc1**
- **ýt$/k5.*mF>yč/k1gMnSc1**
- **á((.|žd)[aeěi])t$/k5.*mF>a$1č/k1gMnSc1**
- **í(.[aeěi])t$/k5.*mF>i$1č/k1gMnSc1**
- **ou(.)it/k5.*mF>u$1ič/k1gMnSc1**
- **í(.)at$/k5.*mF>ě$1ač/k1gMnSc1**
- **ou(.|ch)at$/k5.*mF>u$1ač/k1gMnSc1**
- **ý(.)at$/k5.*mF>y$1ač/k1gMnSc1**
- **ást$/k5.*mF> adeč/k1gMnSc1**
- **ést$/k5.*mF> etač/k1gMnSc1**
- **[éí]ct$/k5.*mF> ekáč/k1gMnSc1**

výsledky

-č

substituční pravidlo	příklad		dvojice	přegenerovávání
t$/k5.*mF>č/k1gMnSc1	*topit/topič*	719	717	2[173]
at$/k5.*mF>áč/k1gMnSc1	*kopat/kopáč*	23	17	6[174]
ít$/k5.*mF>eč/k1gMnSc1	*mlít/mleč*	1	1	0
[eí]t$/k5.*mF>ič/k1gMnSc1	*držet/držič*	5	2	3[175]
ýt$/k5.*mF>yč/k1gMnSc1	*mýt/myč*	1	1	0
á((.|žd)[aeěi])t$/k5.*mF>a$1č/k1gMnSc1	*pálit/palič*	94	92	2[176]
í(.[aeěi])t$/k5.*mF>i$1č/k1gMnSc1	*řídit/řidič*	13	13	0
ou(.)it/k5.*mF>u$1ič/k1gMnSc1	*loupit/lupič*	8	6	2[177]
í(.)at$/k5.*mF>ě$1ač/k1gMnSc1	*vzpírat/vzpěrač*	23	23	0
ou(.|ch)at$/k5.*mF>u$1ač/k1gMnSc1	*foukat/fukač*	10	10	0
ý(.)at$/k5.*mF>y$1ač/k1gMnSc1	*ohýbat/ohybač*	6	6	0
ást$/k5.*mF> adeč/k1gMnSc1	*klást/kladeč*	1	1	0
ést$/k5.*mF> etač/k1gMnSc1	*plést/pletač*	5	5	0
[éí]ct$/k5.*mF> ekáč/k1gMnSc1	*séct/sekáč*	2	2	0
CELKEM	**14 pravidel**	**911**	**896**	**15**
Výjimky[178]	*vozač, trubač*			

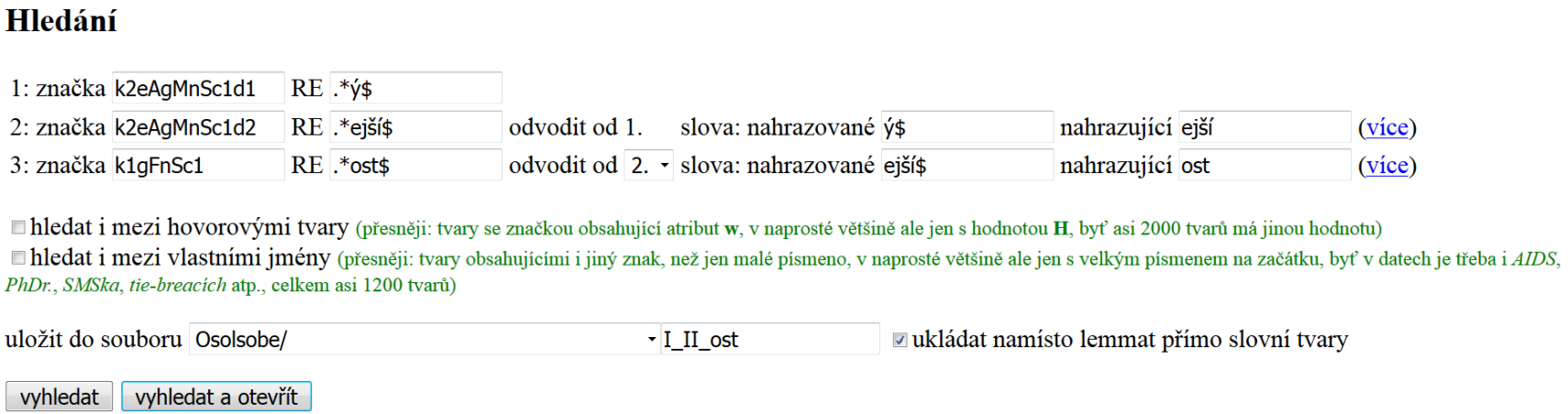
Výhody a nevýhody práce s nástrojem *Deriv*

- Korpus i slovník
- Slovník obsahuje řadu neužívaných výrazů
- Korpus zahrnuje výrazy, které nezná automatický morfologický analyzátor
- I ve slovníku jsou zahrnuty výrazy generované automaticky

Hledání automaticky generovaných tvarů a pozorování frekvencí

- ý$/k2eAgMnSc1d1>ejší/k2eAgMnSc1d2

Hledání trojic



Úkol na příště (28. 10.)

- Zajistit si přístupová práva ke slovníkům pod debdictem
- Zajistit si přístupová práva ke sketchengine
- Podle návodu zpracujte substantiva typu *-ba* (jako *služba, stavba, volba, …*).
- Popište problémy, na které jste při práci narazili a připravte si dotazy k technickým problémům.
- Porovnejte svá zjištění s hesly v SAUČ