CZECH NATIONAL
CORPUS

CZECH NATIONAL CORPUS

Workshop o paralelním korpusu InterCorp

Praha, 6.9.2013

Olga Richterová, ÚČNK

Český národní korpus (LM2011023; 2012-2016)
Ministerstvo školství, mládeže a tělovýchovy
Projekty velkých infrastruktur pro VaVaI

MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY

PŘEHLED PROGRAMU

- 10:00–11:00
  - Typy výzkumných otázek
  - Typy dotazů, regulární výrazy
  - Statistiky (frekvenční distribuce)
- 11:15–12:45
  - Pokročilé dotazy (CQL)
  - Vytváření subkorpusů, podmínky within
  - Kolokace, (třídění)
- 12:45 – Oběd
- 14:15 – Odpolední část programu

Co ne/stihneme

- Ukážeme si témata / cesty / tipy a triky
- Zodpovíme vaše dotazy
- Nestihneme všechna probíraná témata důkladně procvičit
- Tato prezentace bude posléze k dispozici
- V horizontu několika týdnů bude zpřístupněno nové webové rozhraní a posléze nový webový manuál pro práci s korpusy!

Pro a proti korpusového výzkumu

Vzorek jazyka

Data bez interpretace?

Výhody korpusu oproti webu?

- **reprezentativní** – vytvořený z pečlivě vybraných textů
- **neměnný** (referenční) – po zveřejnění se nemění
- **standardizovaný**
- **anotovaný** – opatřený dalšími informacemi

TYPY OTÁZEK

Typy výzkumných otázek

Jaký jazyk nás zajímá?

mluvený, psaný?

publicistiky, beletrie nebo odborné literatury?

překladový, původní?

současný, historický?

Výběr korpusu

Typy výzkumných otázek

Jaký jev chceme hledat?

- konkrétní tvar
  - odvozeniny od základu .*love.*
  - slovo rýmující se s „mírnyx dírnyx“ ([word=".*[iy](x|ks)"])
- všechny tvary jednoho slova
- ustálené slovní spojení/kolokaci
- další informace (z tagů: slovní druh, pád, čas, …)

➢ možnost výzkumu mnoha rovin jazyka

Výběr dotazu

Co nám v současnosti umožňují zdrojová data?

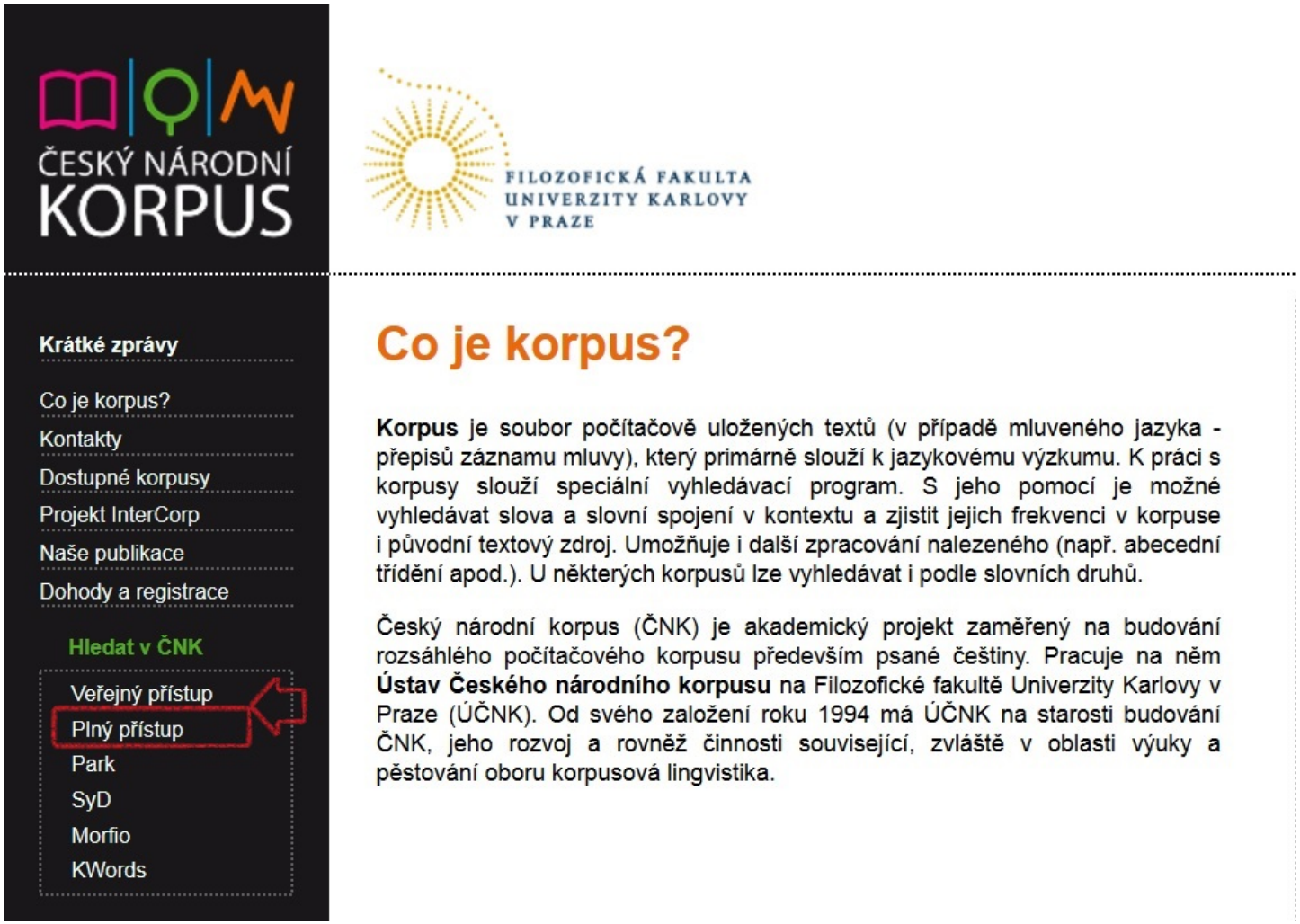
Jaký jazyk InterCorp umožňuje zkoumat?

- psaný, z 20.-21. století
  - publicistický – bez znalosti výchozího textu
  - jazyk beletrie
  - jazyk odborný – právnické texty
- překladový / jazyk originálů

Jevy podle zpracování konkrétního korpusu

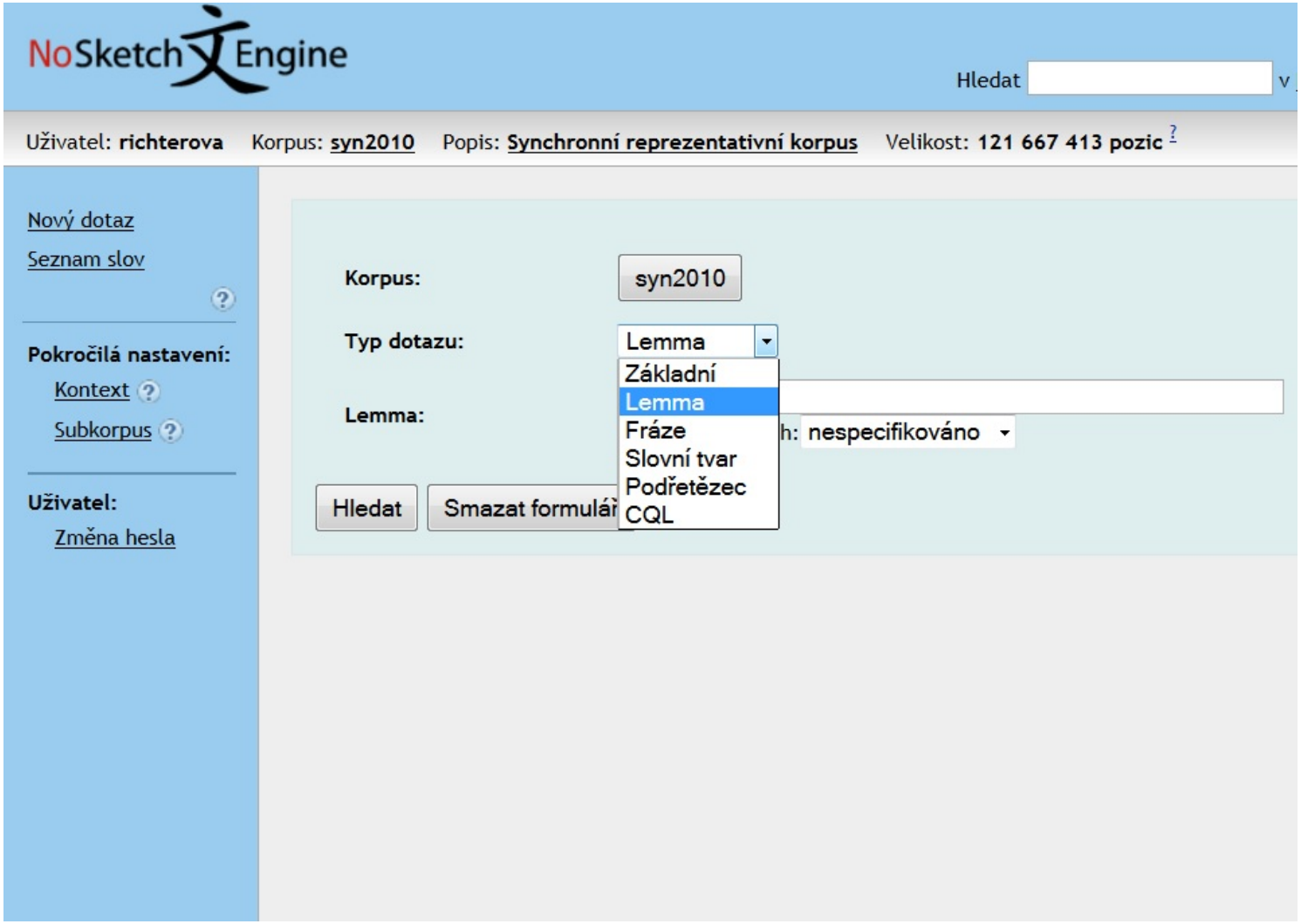
TYPY DOTAZŮ

Začínáme vyhledávat

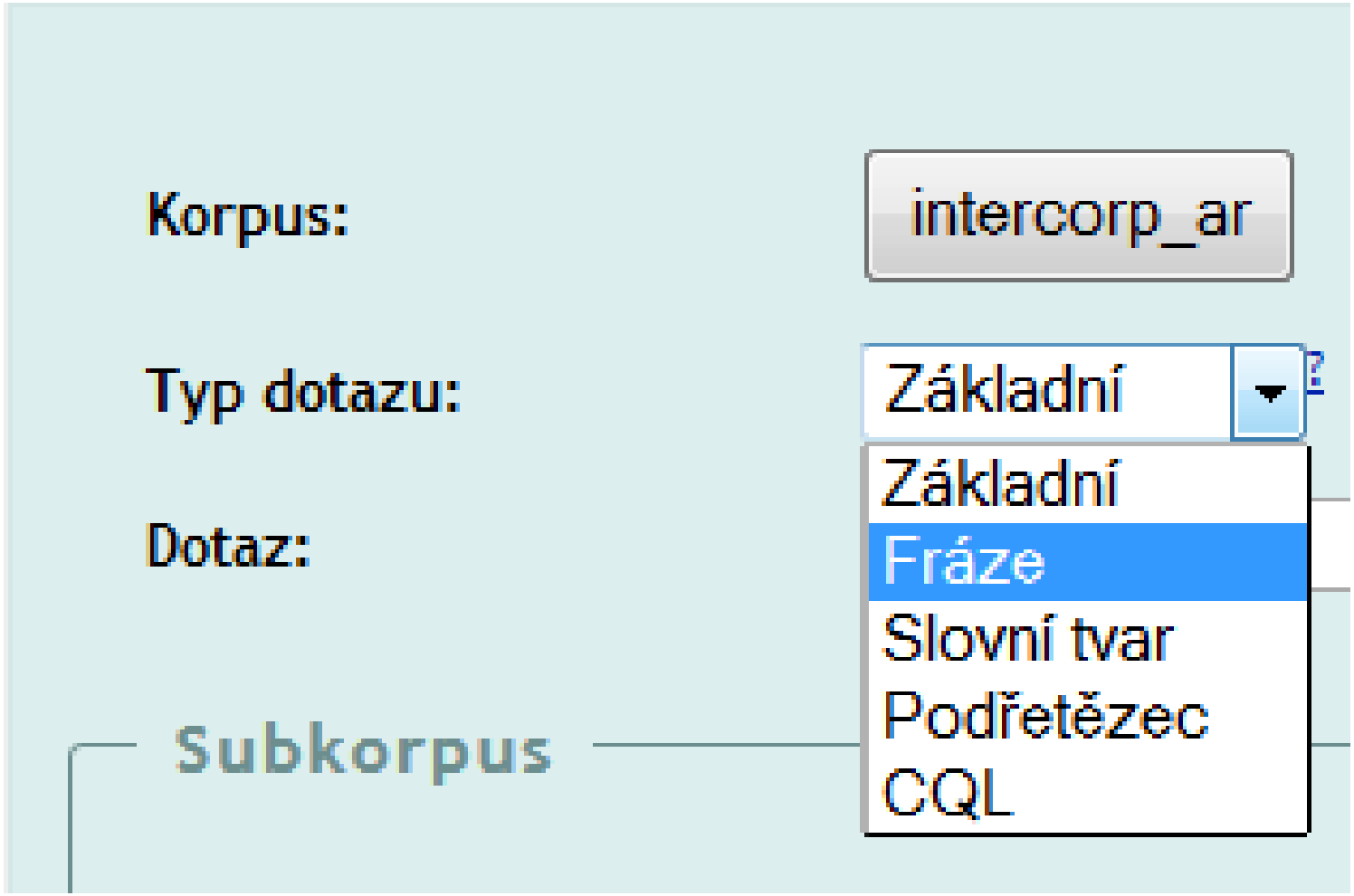


Začínáme vyhledávat

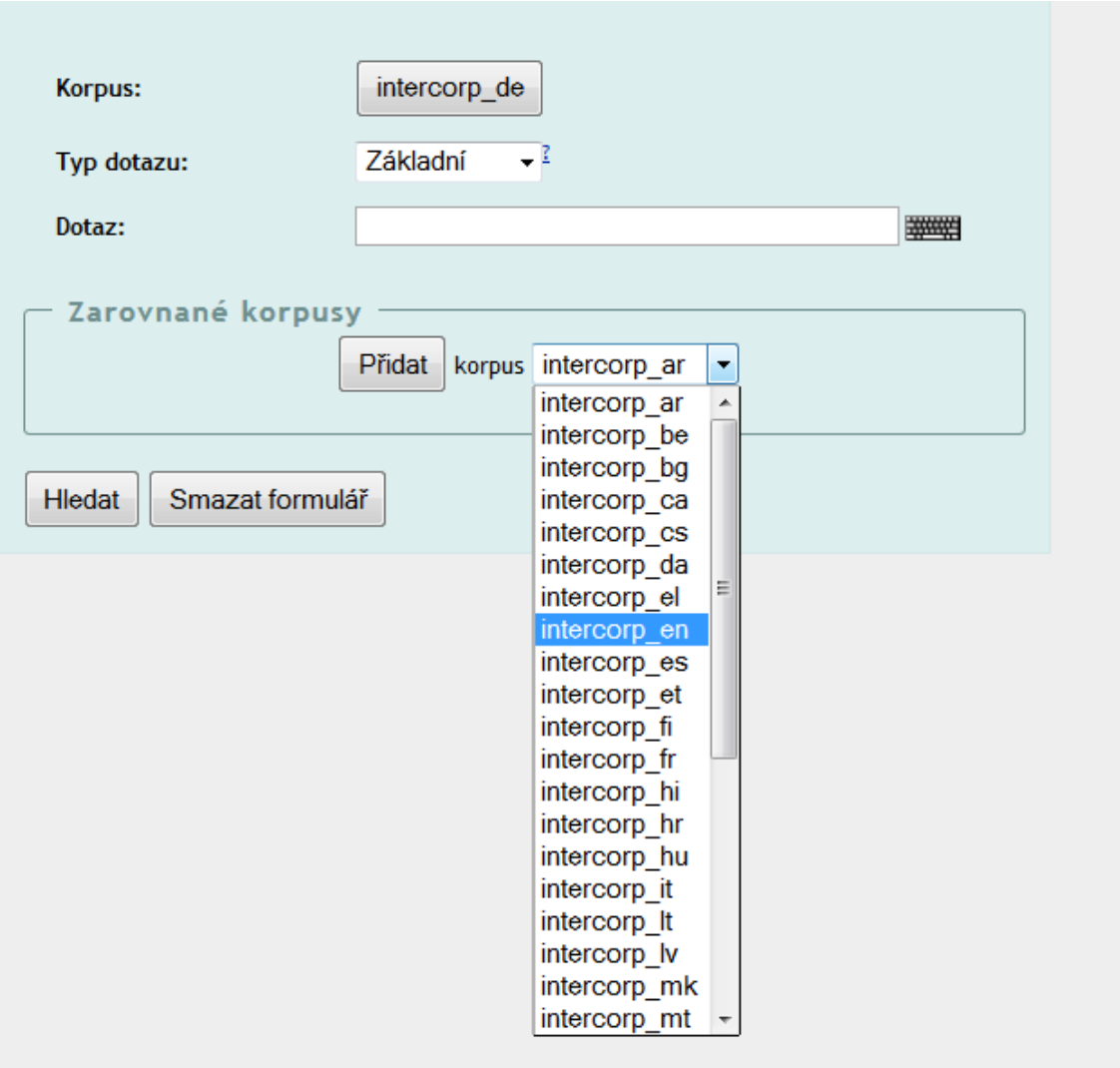
Typy dotazů se liší podle korpusu



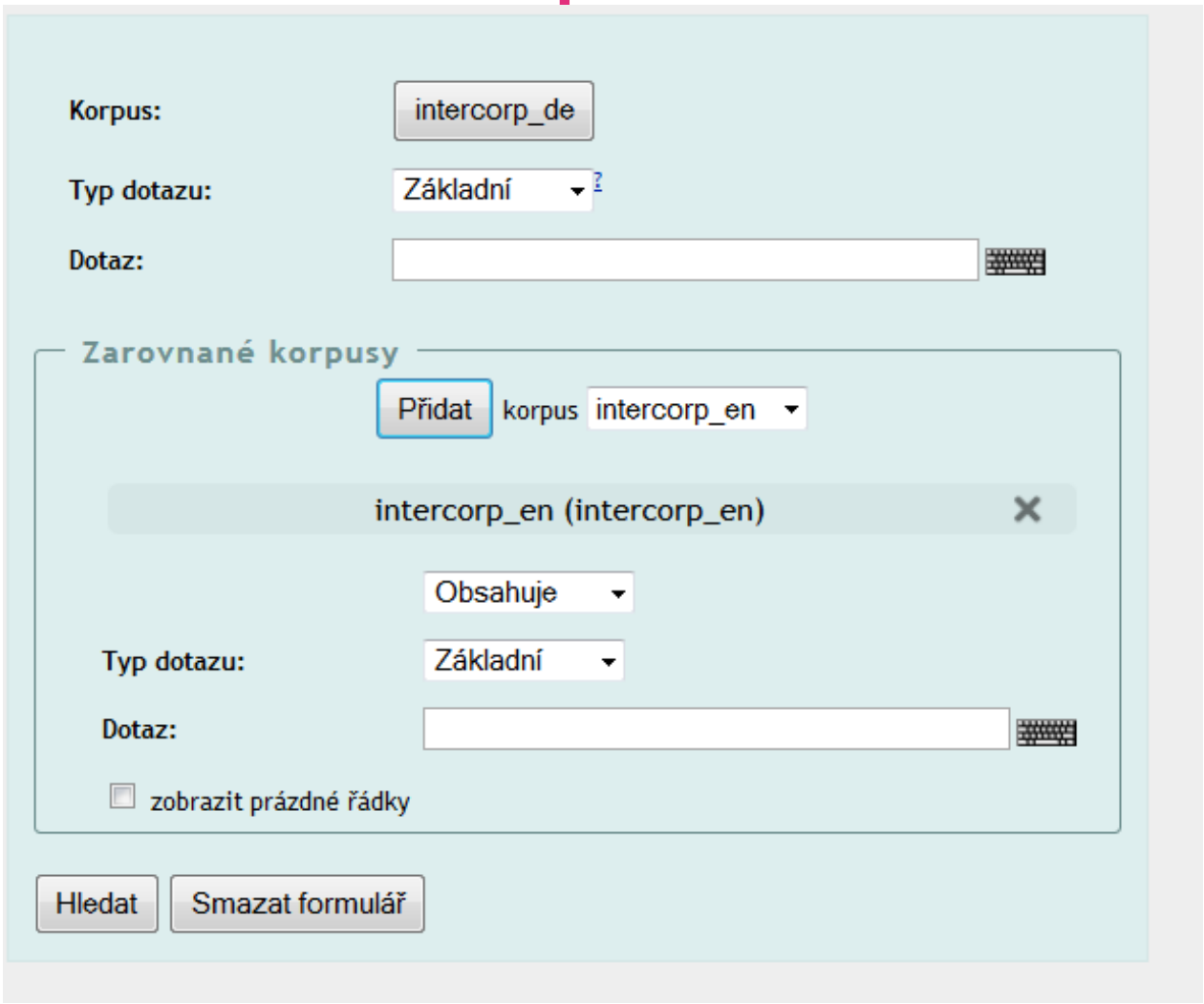
Ne každý je lemmatizovaný…



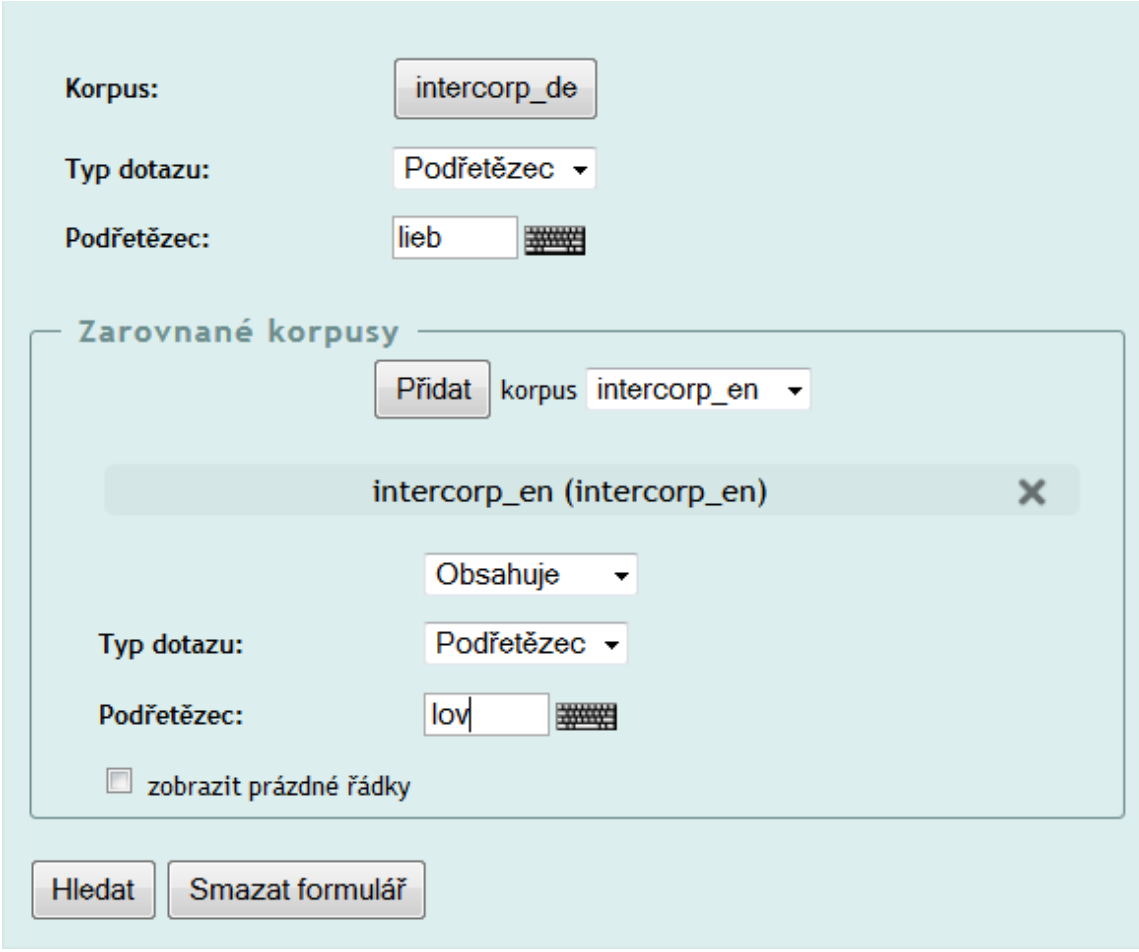
Přidávání paralelních korpusů



Přidávání dalších paralelních korpusů



Paralelní korpus – zadání dotazu



Paralelní – výsledky
typ dotazu: podřetězec

Výskytů: **1 300** i.p.m.: 18,98 (vztaženo k celému korpusu) | ARF: **269**

strana 1 ze 65 Přejít další | poslední

	intercorp_de		intercorp_en
Per Anhalter durch die Galaxis	Zaphod liebte die große Show : darin war er der Größte .	The Hitch Hiker's Guide to the Galaxy	Zaphod loved effect : it was what he was best at .
Per Anhalter durch die Galaxis	Das passte den Dentrassis ausgezeichnet in den Kram , denn sie liebten das vogonische Geld , das eine der härtesten Währungen im Weltall ist , aber die Vogonen konnten sie nicht leiden .	The Hitch Hiker's Guide to the Galaxy	This suited the Dentrassis fine , because they loved Vogon money , which is one of the hardest currencies in space , but loathed the Vogons themselves .
Per Anhalter durch die Galaxis	Zu Hause war diese Industrie auf dem Planeten Magrathea , wo Hyperraum-Ingenieure durch weiße Löcher im All Materie ansaugten und sie in liebevoll gestaltete Traumplaneten verwandelten - Goldplaneten , Platinplaneten , Weichgummiplaneten mit Massen von Erdbeben - , die selbst den verwöhntesten Ansprüchen der reichsten Männer der Galaxis genügten .	The Hitch Hiker's Guide to the Galaxy	The home of this industry was the planet Magrathea , where hyperspatial engineers sucked matter through white holes in space to form it into dream planets - gold planets , platinum planets , soft rubber planets with lots of earthquakes - all lovingly made to meet the exacting standards that the Galaxy 's richest men naturally came to expect .
Die Brücke über die Drina	Das verdirbt das Spiel und ruft Enttäuschung und Unmut bei denen hervor , die das Spiel der Phantasie lieben , die Ironie hassen und glauben , mit geduldigem Hinschauen könne man wirklich etwas sehen und erleben .	The bridge on the Drina	Openmouthed they would peer into that deep dark hole , quivering with curiosity and fear , until it seemed to some anaemic child that the opening began to sway and to move like a black curtain , or until one of them , mocking and inconsiderate (there is always at least one such) , shouted ' The Arab ' and pretended to run away . That spoilt the game and aroused disillusion and indignation amongst those who loved the play of imagination , hated irony and believed that by looking intently they could actually see and feel something .
Die Brücke über die Drina	Und als sie dann am nächsten Tage , am trüben Morgen , vom Hügel auf diese Stadt hinunterblickten , die sie unbewußt und stark wie ihr eigenes Herzblut liebten , und das trübe hochgehende Wasser betrachteten , wie es reißend in Höhe der Hausdächer durch die Straßen strömte , dann errieten sie an diesen Dächern , von denen das Wasser mit Krachen Brett um Brett losriß , wessen Haus noch stand .	The bridge on the Drina	When the next day , in the cloudy dawn , they looked down from the hillside on the town that they loved as strongly and as unconsciously as their own blood , and saw the darkened muddied waters rushing through the streets at roof level , they would try to guess whose house it was from which the foaming waters were noisily tearing the roof plank by plank and whose house still remained upright .

Závislost typu dotazu na korpusu

V lemmatizovaných korpusech je možné hledat

- konkrétní, použitý slovní tvar (**word**) – např. *kočce, běž, gelaufen, headings*
- základní slovníkový tvar (**lemma**) – např. *kočka, běžet, laufen, heading*

V označkovaných (otagovaných) korpusech lze najít i morfologickou značku (**tag**)

V anotovaných korpusech lze zadat i další podmínky

Typy dotazů

typ dotazu	s / bez RE (regulárních výrazů)	počet slov	další
základní	bez	více	zadáme-li tvar lemmatu, vyhledá celé paradigma
lemma	s	jen 1	lze specifikovat sl. druh (*stát* jako sloveso)
fráze	s	více	konkrétní slovní tvary
slovní tvar	s	jen 1	lze specifikovat sl. druh (*při* jako podst. jm. – od *pře*)

Typy dotazů – dokončení

typ dotazu	s / bez RE (regulárních výrazů)	počet slov	další
podřetězec	s	1 řetězec	vyhledá např. *mrsk* – všechny odvozeniny slov *mrskat, mrsknout, smrsknout, Zámrsk*, i překlepy typu *mrskev*
CQL	s – umožňuje nejpřesnější dotazování a kombinaci různých kritérií	více	umožní zadat podmínky a dotázat se na libovolný počet pozic

Co jsou to regulární výrazy

Regulární výrazy: zástupné symboly a možnosti opakování

- Mohou se užívat ve všech typech dotazů kromě základního
  - **tečka** (.) – představuje jeden libovolný znak,
  - **interval** ({n, k}) – *n* až *k* opakování předchozího znaku nebo většího celku,
  - **hvězdička** (*) – libovolný počet (0 a více) opakování předchozího znaku nebo celku, tj. {0,}
  - **plus** (+) – 1 nebo více opakování předchozího znaku nebo celku, tj. {1,}

Regulární výrazy: možnosti opakování a logické operátory

- **otazník** (?) – žádný nebo jeden výskyt předchozího znaku nebo celku, tj. {0,1}
- **seznam** ([]) – alternativa, výběr jednoho libovolného znaku z těch, které jsou uvedeny uvnitř závorek
- **svislá čára** (|) – také alternativa, ne ovšem mezi jednotlivými znaky, ale celými řetězci tvořícími jednotku
- **kulaté závorky** – libovolnou část výrazu je možné seskupit do kulatých závorek, vytvořit tak jistý celek a ovlivnit tím prioritu jeho vyhodnocování

Regulární výrazy a dotazovací jazyk

Více informací k regulárním výrazům:

https://www.korpus.cz/bonito/regular.php

Více informací k dotazovacímu jazyku (anglicky)

http://trac.sketchengine.co.uk/wiki/SkE/CorpusQuerying

Vnitřní struktura korpusu

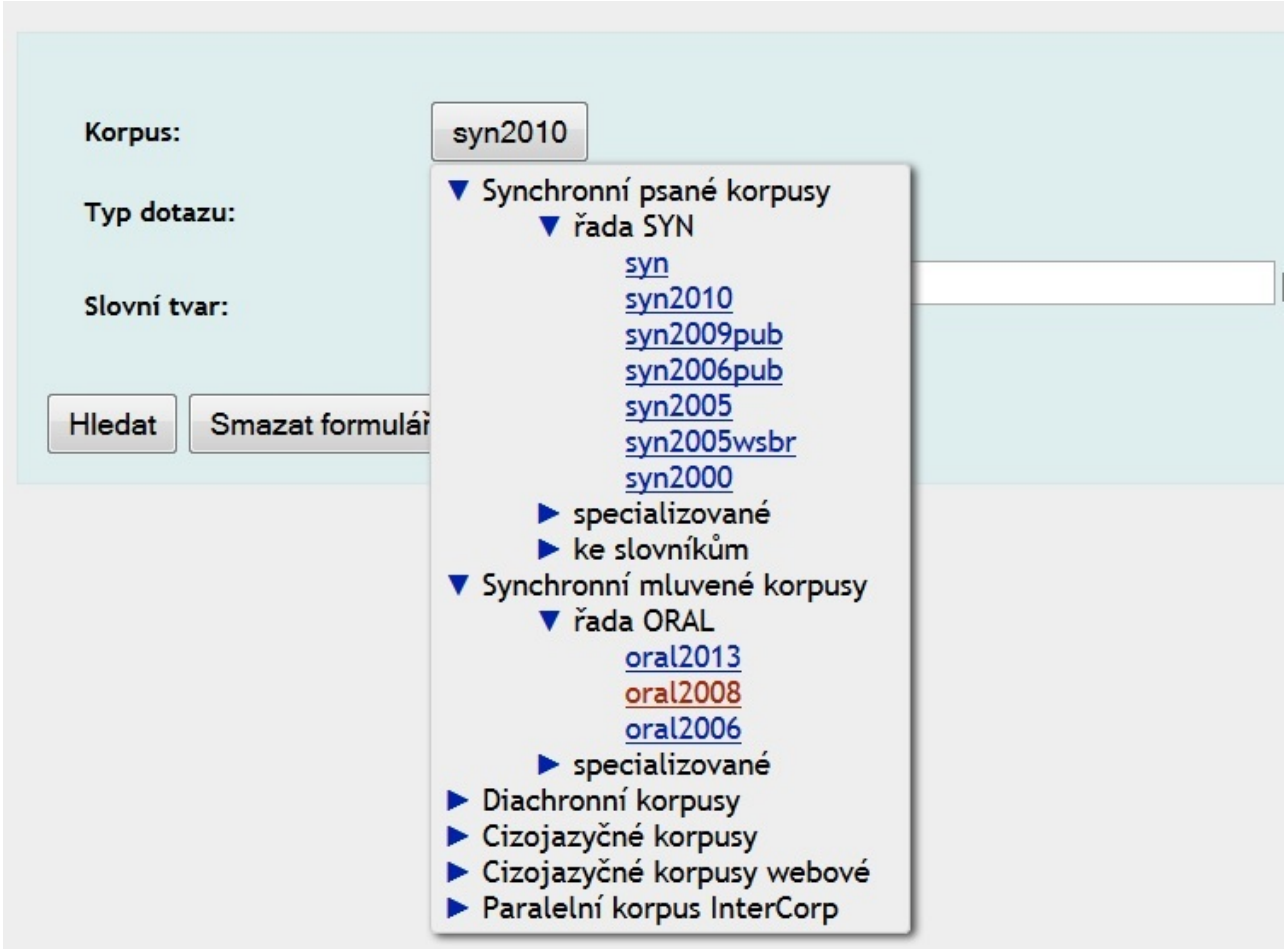
- Zjednodušené uspořádání dat v lemmatizovaném a tagovaném korpusu:

slovní tvar (word)	lemma	tag (zkrácený)
Když	když	J.*
školení	školení	N..S4.*
skončilo	skončit	V.*
,	,	Z.*
… <s/>		

Základní dotazy v novém rozhraní
korpus SYN2010

- **Základní dotaz**: vyhledejte *prašivý pes* a *černá kočka.* V čem se liší výsledky?
- **Lemma:** vyhledejte
    - … (tři tečky)
    - .+nést
    - ra(ta)+
    - ps*t
- **Slovní tvar:** vyhledejte
    - … (tři tečky)
    - při (a specifikujte slovní druh jako podstatné jméno)

Změna vybraného korpusu



Základní dotazy: nelemmatizovaný korpus

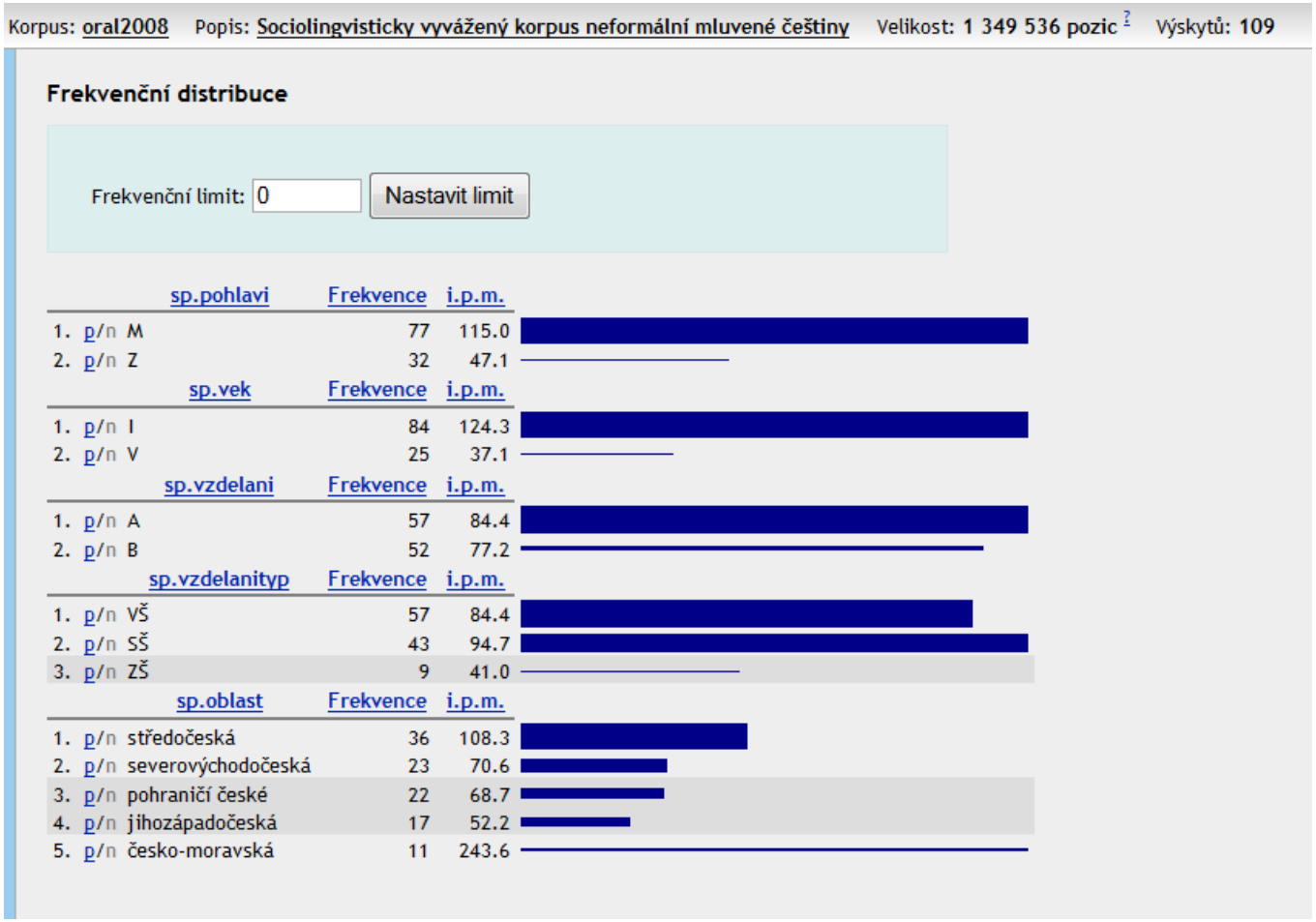
- Vyhledejte v korpusu ORAL2008
  - v typu dotazu *Základní*
    - a?[nj]o
  - v typu dotazu *Slovní tvar*
    - a?[nj]o
    - tuhle.+
    - .*(dle|hle)nc.*
    - .*[dh]lenc.*

Shrnutí typů dotazů

- Existují různé typy dotazů, které většinou umožňují využívat zástupné symboly (tzv. regulární výrazy)
- Regulární výrazy nabízejí mnohem širší vyhledávací možnosti než pouhé řetězce písmen
- Nejpřesnější pokládání dotazů umožňuje dotazovací jazyk CQL

STATISTIKY / Frekvenční distribuce

Frekvenční distribuce – ORAL2008



Korpus: oral2008 Popis: Sociolingvisticky vyvážený korpus neformální mluvené češtiny Velikost: 1 349 536 pozic Výskytů: 109

Frekvenční distribuce

Frekvenční limit: 0 Nastavit limit

	sp.pohlavi	Frekvence	i.p.m.
1.	p/n M	77	115.0
2.	p/n Z	32	47.1
	sp.vek	**Frekvence**	**i.p.m.**
1.	p/n I	84	124.3
2.	p/n V	25	37.1
	sp.vzdelani	**Frekvence**	**i.p.m.**
1.	p/n A	57	84.4
2.	p/n B	52	77.2
	sp.vzdelanityp	**Frekvence**	**i.p.m.**
1.	p/n VŠ	57	84.4
2.	p/n SŠ	43	94.7
3.	p/n ZŠ	9	41.0
	sp.oblast	**Frekvence**	**i.p.m.**
1.	p/n středočeská	36	108.3
2.	p/n severovýchodočeská	23	70.6
3.	p/n pohraničí české	22	68.7
4.	p/n jihozápadočeská	17	52.2
5.	p/n česko-moravská	11	243.6

Frekvenční distribuce I

- Distribuce lemmat a zdrojový jazyk
  - InterCorp EN: [lemma=„s?he“]

lemma	absolutní frekvence	podíl
he	177 205	68,7 %
she	80 663	31,3 %
celkem	**257 868**	**100 %**

  - InterCorp EN: [lemma=„s?he“], srclang=„en“

lemma	absolutní frekvence	podíl
he	95 875	65,8%
she	52 148	34,2 %
celkem	**148 023**	**100 %**

Frekvenční distribuce II

- Porovnání s angličtinou
  - SYN (CS): [lemma="ona?"]
  - InterCorp CS: [lemma="ona?"]

  ➢Nelze, v češtině *ona* lemmatizováno jako *on*

  ➢*Nebo přece...?* Ukážeme si později.

  - Je tomu stejně i u přivlastňovacích zájmen?

  ➢Zkusme se zeptat na [lemma=„je(ho|jí)“]

Frekvenční distribuce III

- Projevuje se v distribuci přivlastňovacích zájmen v CS a EN typologický rozdíl mezi jazyky?
  - Rozhodně je vidět rozdíl v tagování!

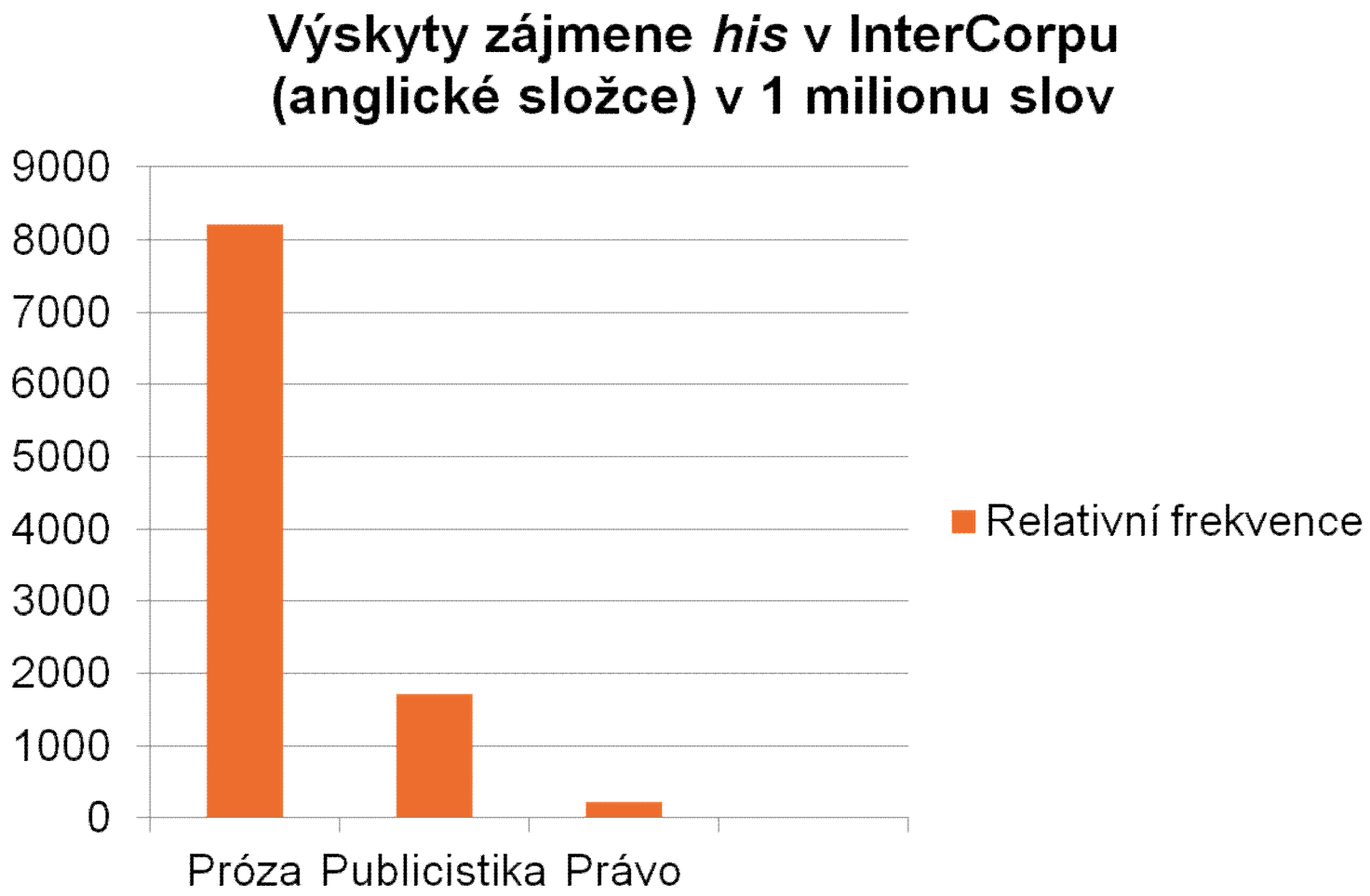
lemma	korpus	abs. frekvence	relat. frekvence
his (DPS, PNP, UNC) (PP$)	BNC InterCorp	409 825 117 914	**3 684** **1 826**
originál: *his*	**InterCorp EN**	**54 685**	**7 774**
her (DPC, UNC, VVG) (PP$, PP)	BNC InterCorp	23 376 77 410	**1 828** **1 199**
jeho	SYN	2 601 136	**1 658**
její	SYN	1 163 534	**742**

Pozor na skladbu korpusů

	Text type	Frekvence	i.p.m.
1.	p/n Written books and periodicals	375 868	4185.4
2.	p/n Written miscellaneous	13 916	1685.8
3.	p/n Spoken context-governed	7 304	1066.3
4.	p/n Spoken demographic	6 762	1382.6
5.	p/n Written-to-be-spoken	5 975	4137.4
	Publication date	**Frekvence**	**i.p.m.**
1.	p/n 1985-1993	369 906	3636.1
2.	p/n 1975-1984	22 460	4157.9
3.	p/n 1960-1974	13 697	6661.7
4.	p/n Unknown	3 762	1830.4
	Domain for written corpus texts	**Frekvence**	**i.p.m.**
1.	p/n Imaginative	155 523	7906.2
2.	p/n Informative: world affairs	83 248	4327.6
3.	p/n Informative: leisure	52 643	3837.9
4.	p/n Informative: arts	38 146	5105.1
5.	p/n Informative: social science	27 632	1764.9
6.	p/n Informative: belief & thought	15 043	4426.4
7.	p/n N/A	14 575	1237.6
8.	p/n Informative: commerce & finance	12 478	1537.7
9.	p/n Informative: applied science	7 506	944.6
10.	p/n Informative: natural & pure science	3 031	712.3

- Rozdíl mezi jazyky X rozdíl mezi text. typy/žánry?
  - BNC a InterCorp jsou sestavené jinak!
  - Nejvyšší frekvence v BNC: Imaginative: 7906 i.p.m.

Rozdílná distribuce v textových typech



POKROČILÉ DOTAZY

Dotazovací jazyk a uplatnění více podmínek na tutéž pozici (slovo)

- CQL (corpus query language): [atribut="hodnota"]
- logické operátory: & (zároveň), | (nebo), ! (negace)
  - SYN2010: [lemma="on" & tag="P.F.*"] (tj. lemma "ona")
    - 2 867 i.p.m., nelze v InterCorp (CS)
  - Intercorp (EN) – homonymní tvary (*states:* tag N.*|V.*)
    - [word="states"&tag!="N.*"]
      - 1 608 výskytů, frekv. distrib. – typy textu: seřadit dle i.p.m.
        - EuroParl – 42 i.p.m. (výskytů na milion)
        - publicistika – zprávy – 34 i.p.m.
        - próza – 2 i.p.m.

Dotazovací jazyk a prázdná pozice

[lemma="have"][][lemma="get"]

Výskytů: **1 114** i.p.m.: 17,25 (vztaženo k celému korpusu) | ARF: **281** | Výsledek je promíchán

strana 1 ze 56 Přejít další | poslední

	intercorp_en		intercorp_cs
_EUROPARL,NE,de	We have to get involved through education .	_EUROPARL,NE,de	Musíme se zabývat vzděláváním .
_PRESSEUROP,cs	A month after general elections , the Czech Republic has now got a new prime minister .	_PRESSEUROP,cs	Měsíc po legislativních volbách , Česká republika má nového premiéra .
_SYNDICATE	How much will they have to get angry about then because of the way that we are behaving today ?	_SYNDICATE	Kolik důvodů ke vzteku budou v té době mít kvůli tomu , jak se chováme dnes ?
munro-utek,ANO,en	She had to get herself in gear to face the day , working in the newspaper office .	munro-utek,NE	Musí se dostat do formy , aby zvládla všechno , co ji čeká v redakci .
_EUROPARL,NE,fr	Now we are going to have to get down to the practical matters and , in particular , to an essential issue that we have been skirting round , that of own resources .	_EUROPARL,NE,fr	Nyní se budeme muset začít věnovat praktickým záležitostem a jmenovitě jistému zásadnímu tématu , kterému se vyhýbáme , tj . tématu vlastních zdrojů .
frost-sez_sedmi,ANO,en	" Come on , Doyle , we have n't got all night , " he said .	frost-sez_sedmi,NE,en	" Honem , Doyle , nemáme na to celou noc , " naléhal .
day-cirkus_v_zime,ANO,en	She remembered the atomic bombs , the first , the second , but for the third , Ethan had n't gotten up to get his wallet , and she 'd been too busy to notice .	day-cirkus_v_zime,NE,en	Uvědomila si , že před výbuchem té třetí už Ethan nevstal , nic si z peněženky nevzal a ona měla v hlavě jiné věci , takže to ani pořádně nezaznamenala .
styblova-skalpel_pros,NE,cs	We got him two buttered rolls , which he scoffed before we had even got out of Prague .	styblova-skalpel_pros,ANO,cs	Sehnali mu dvě housky s máslem a on je zbaštil , ještě než jsme vyjeli z Prahy .
chevalier-divka_s_perl,NE	' You will have to get one of your own some day , ' I heard him say in his deep voice .	chevalier-divka_s_perl,NE	„ Jednoho dne si budete muset pořídit svoji vlastní , " slyšela jsem ho , jak říká svým hlubokým hlasem .
Amis-Stastny_Jim,ANO,en	The Margaret business has been getting me down rather . '	Amis-Stastny_Jim,NE,en	Celá ta záležitost s Margaretou mi začíná lézt na mozek . "
mulisch-atentat,NE,nl	" Ca n't you imagine that we had to get out of there in a hurry , after the War ? "	mulisch-atentat,NE,nl	" Nenapadlo tě , že jsme hned po válce museli koukat mazat ? "
styblova-skalpel_pros,NE,cs	They have to get order into their blood , there ' s no room for slackness .	styblova-skalpel_pros,ANO,cs	Někdy je honím pro hloupost , jinak to nejde , musí dostat pořádek do krve .
_EUROPARL,NE	It has never got out of it .	_EUROPARL,NE	Tato země se z ní nikdy nevzpamatovala .
Mandelstamova-DveKnihy,NE,ru	She soon stopped coming , and the student who had pressed her on me , a decent girl who had obviously got caught in the web , was clearly very upset and anxious to explain things to me .	Mandelstamova-DveKnihy,NE,ru	Po Larisině odhalení práskačka rychle zmizela , zato studentka , která mi ji doporučila , hodná dívenka , jež nepochybně sama naletěla , prožívala celou záležitost jako velkou tragédii a pořád se mi snažila něco vysvětlovat .
_EUROPARL,NE	Having said that , I would also echo what was said by a large number of Members on the fact that in order to get price stability we had to get some cooperation from other decision-makers , authorities and the private sector .	_EUROPARL,NE	Po této poznámce bych chtěl reagovat na slova mnoha poslanců o tom , že k získání cenové stability musíme získat určitou spolupráci ostatních subjektů s rozhodovací pravomocí , úřadů a soukromého sektoru .
asimov-ocelove_jesky,ANO,en	There is n't one day , not one damned hour , that we have n't got cultures of every strain of yeast in the company growing in our kettles .	asimov-ocelove_jesky,NE,en	Nemine den , ani jediná pitomá hodina , kdy bychom nevypěstovali v našich pokusných konvích smíšené kultury všech kvasinkových rodů , kontrolujeme je a upravujeme , aby kryly potřebu potravin .

Dotazovací jazyk a prázdná pozice

- CQL a tokenizace

	Cokoli	Určená pozice	Určená pozice	Určená pozice
Dotaz		**[lemma="have"]**	**[]**	**[lemma="get"]**
Realizace	we	have	n't	got
Realizace	She	had	to	get
Realizace	Republic	has	now	got

Dotazovací jazyk a operátor rozsahu

- Tento dotaz:
  - [lemma="have"][][lemma="get"]
- se rovná:
  - [lemma="have"][]{1}[lemma="get"]

- Zkusme najít věty tázací:
  - [lemma="have"][]{1,2}[lemma="get"][]+[word="\?"]
  - ➢ omezí se vyhledávání na 1 větu?

Dotazovací jazyk

	intercorp_en		intercorp_cs
brown-chut_lasky,ANO,en	Have you got balls enough to face me like a man ? Or are we gonna continue this silly game of hide-and-seek ? " Following a short silence , a voice came to him from the other side of the wall . " Threadgill ? "	brown-chut_lasky,NE,en	Máš dost odvahy , aby ses mi postavil jako chlap ?
styblova-skalpel_pros,NE,cs	How had it got here ? What ? Mr Materna gave it to us ?	styblova-skalpel_pros,ANO,cs	Jak se to tu octlo ?
munro-utek,ANO,en	How could she be sure that they had not got her as a replacement ? If there was one big thing she hadn ' t known about , why could there not be another ?	munro-utek,NE	Cožpak věděla s jistotou , že si ji nepořídili jako náhradu ?
Brown-zdravim_temnoto,ANO,en	Spotting a group of acquaintances beyond Gavin 's shoulder , she waved , calling out , " Hey , y'all , I 'm back from France , and have I got stories ! " Gavin sidestepped , blocking her view of the others and forcing her to look at him . " Is Janey really missing ? "	Brown-zdravim_temnoto,NE,en	Gavinovi přes rameno uviděla hlouček nějakých známých , zamávala na ně a zavolala : „ Ahoj , zrovna jsem přijela z Francie , počkejte , až vám budu vyprávět ! “
woolfova-dallowayova,ANO,en	What had she got in her work-box ?	woolfova-dallowayova,NE,en	Co ještě má v košíčku na šití ?
kohout-snezim,NE,cs	I 'll lose , of course , but first I have to get some guarantee from her that will reassure me . " It 's , like , that guy who came to see me . He brought me back , too , so I could tell you . " Liar ! You needed clean pants ; at least there 's that ! " What kind of guy is he ? "	kohout-snezim,ANO,cs	(Prohraju , ovšemže , ale předtím z ní musím dostat záruky , které mě aspoň trochu uklidní .)
adams-stoparuv_pruvodc,ANO,en	" Well , I 've just got all these bulldozers and things to lie in front of because they 'll knock my house down if I do n't , but other than that ... well , no not especially , why ? " They do n't have sarcasm on Betelgeuse , and Ford Prefect often failed to notice it unless he was concentrating . He said , " Good , is there	adams-stoparuv_pruvodc,NE,en	" No , jenom musím ležet před všemi tady těmi buldozery , nebo co to je , protože , když to neudělám , tak mi zbourají dům , ale jinak ... vlastně ani ne , proč ? "

PODMÍNKY A SUBKORPUSY

Podmínky: v rámci jedné věty

- [lemma="have"][]{1,2}[lemma="get"][]+[word="\?"] within <s/>

What	've you got there ?	VÉNA har
said Thorin , "	have n't you got a map ?	and did n't you hea
ack . "	Has n't anyone got any sense ?	We 've got to relig
id heat . When	had it gotten so hot ?	I did not remembe
p hiss . " What	has it got in its pocketses ?	Tell us that . It
me , Mr Dixon ;	have you got a minute to spare ?	' First ma
e said . " What	have you got that you ' re ashamed of ?	Have you a smelly
' And what else	have you got me for ?	" " You '
/p> " When	have you got to go ?	" he asked in an
> LUCY : What	have you got to do ?	(She bur
> " Lozada !	Have you got balls enough to face me like a man ?	Or are we gonna c
/p> " What	have you got there ?	" she asked curiou
it ; exactly how	had she got him to come inside her room ?	And did n't she ha
' What	have you got in the basket , Pippi ?	' asked Annika . '
ment how they had come to get involved with him - he was obviously already suffering from sclerosis and softening of the brain , so what would he be like later ?		M. explained that
' No one-fifty .	Have n't got it mixed up with the one-forty , by any chance ?	' Dixon sv
I 'm back now .	Have you got a job for me ?	" Armans
n countries , or	have I got the wrong person ?	Now , Eu
kson ?	Have you got a pet animal ?	Will you g
too late . What	has it got in its pocketses ?	he cried . The ligh

Vyhledávání dle větné pozice

Adverbiale na počátku věty v angličtině

- Chceme najít krátká příslovečná určení a jiná uvození v iniciální pozici anglických vět, oddělená čárkou (a vyloučit slovesné tvary):

<s> [word!="V."]{1,2}[word="\,"]

- Stejně tak můžeme vyhledávat např. podstatná jména předcházející konci věty: [tag="N.*"] [] <s/>

Frekvenční distribuce a podmínky

- hledání interjekcí v jazyce konkrétního autora:
- [tag="I.*"] within
    - Jak zjistíme tag anglických interjekcí? – nějakou zadáme a *Frekv. distr > značky*
    - *hey* – UH, NP, *bump* – NN, VB, NP, VBP
- podobně: osobní zájmena v jazyce V. Woolf
    - [lemma="s?he"] within
- pozor: i.p.m. (80 výskytů) vztaženo k celému korpusu!

Vytváření subkorpusů

- ***Subkorpus – Vytvořit nový – Vlastní within podmínka***
  - within
    - 186 222 tokenů
  - Hledat v: *Dostupné subkorpusy*
    - lemma *she* – 17 264 i.p.m.
    - lemma *he* – 12 066 i.p.m.

Rozdíly: SYN(…) a InterCorp

- SYN: není-li u atributu „srclang“, (source language, zdrojový jazyk), uvedena žádná hodnota, jedná se o češtinu.
- Subkorpus obsahující pouze **původně české, nepřekladové texty?**
  - SYN(…): within <srclang=""/>
  - InterCorp: within <srclang="CS"/>

Další rozdíly: SYN(…) a InterCorp

- strukturní atributy:
  - doc – opus – div
- autor – author
- velká/malá písmena u zdroj. jazyka a jejich počet
  - en – ENG
- ...

SYN, InterCorp a jazyk překladů

- SYN: po vytvoření subkorpusu s podmínkou zdrojového jazyka češtiny:
  - [tag="I.*"]within <opus srclang="" />
  - můžeme porovnat např. s citoslovci v jazyce překladu:
  - [tag="I.*"]within <opus srclang!="" />
- Podobně v InterCorpu musíme dbát na směr překladu: u řady textů však neznáme zdrojový jazyk!

Tip: pozor na tagování

- Z rakouské němčiny přejatý výraz pro *rychle*:
  - [lemma=„kách“] (SYN2010, SYN)
  - nalezneme např.: Ti druzí umřeli moc **kách** .
  - Ale také nalezneme *kách* jako koncovku:
  - Určete, ve které (ých) zkumavce (**kách**) vznikla sraženina!

- Proto nás zajímá *kách* jako adjektivum / adverbium:
  - [tag="[AD].*"&word="kách"] – tytéž výsledky

KOLOKACE

Kolokace

- důležitost parametrizace a volby konkrétní míry
- přímý p/n filtr

1. [lemma="nechat"], kolokace v pravém okolí (1-3 pozice)

   rozdíly v uspořádání podle:

   – *MI*: části frazémů a málo frekventované infinitivy
   – *T-score*: gramatická slova
   – *logDice*: něco „mezi“ oběma extrémy

2. odlišný kontext: *statečný* vs. *odvážný*

Děkujeme za pozornost!

olga.richterova@ff.cuni.cz
michal.kren@ff.cuni.cz

CZECH NATIONAL CORPUS