



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 VaVal



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 – právníké texty
- překladový / jazyk originálů



Jevy podle zpracování konkrétního korpusu





# TYPY DOTAZŮ



# Začínáme vyhledávat



ČESKÝ NÁRODNÍ  
KORPUS

---

Krátké zprávy

---

Co je korpus?

---

Kontakty

---

Dostupné korpusy

---

Projekt InterCorp

---

Naše publikace

---

Dohody a registrace

---

Hledat v ČNK

- Veřejný přístup
- Plný přístup
- Park
- SyD
- Morfio
- KWords



## Co je korpus?

**Korpus** je soubor počítačově uložených textů (v případě mluveného jazyka - přepisů záznamu mluvy), který primárně slouží k jazykovému výzkumu. K práci s korpusy slouží speciální vyhledávací program. S jeho pomocí je možné vyhledávat slova a slovní spojení v kontextu a zjistit jejich frekvenci v korpuse i původní textový zdroj. Umožňuje i další zpracování nalezeného (např. abecední třídění apod.). U některých korpusů lze vyhledávat i podle slovních druhů.

Český národní korpus (ČNK) je akademický projekt zaměřený na budování rozsáhlého počítačového korpusu především psané češtiny. Pracuje na něm **Ústav Českého národního korpusu** na Filozofické fakultě Univerzity Karlovy v Praze (ÚČNK). Od svého založení roku 1994 má ÚČNK na starosti budování ČNK, jeho rozvoj a rovněž činnosti související, zvláště v oblasti výuky a pěstování oboru korpusová lingvistika.



# Začínáme vyhledávat



ČESKÝ NÁRODNÍ  
KORPUS

---

Krátké zprávy

Co je korpus?

Kontakty

Dostupné korpusy

Projekt InterCorp

Naše publikace

Dohody a registrace

**Hledat v ČNK**

Veřejný přístup

Plný přístup



## Plný přístup ke korpusům ČNK

Přístup pouze pro registrované uživatele.



Plný přístup ke korpusům ČNK je možné získat na základě souhlasu s Prohlášením uživatele že přístup ke korpusům nepoužije ke komerčním účelům. Prohlášení je ke stažení ve formátu možné ho doručit osobně, zaslat poštou, případně pro vyplnění a odeslání použít elektronick



# Typy dotazů se liší podle korpusu

The screenshot displays the NoSketch Engine search interface. At the top left is the logo 'NoSketch Engine'. To the right is a search bar with the text 'Hledat' and a search icon. Below the logo, the user information is shown: 'Uživatel: richterova', 'Korpus: syn2010', 'Popis: Synchronní reprezentativní korpus', and 'Velikost: 121 667 413 pozic'. On the left side, there is a navigation menu with links for 'Nový dotaz', 'Seznam slov', 'Pokročilá nastavení' (with sub-links for 'Kontext' and 'Subkorpus'), and 'Uživatel' (with a link for 'Změna hesla'). The main search area contains a form with the following fields: 'Korpus:' with a text input containing 'syn2010'; 'Typ dotazu:' with a dropdown menu currently showing 'Lemma' and a list of options including 'Základní', 'Lemma', 'Fráze', 'Slovní tvar', 'Podřetězec', and 'CQL'; and 'Lemma:' with a text input field. Below the 'Typ dotazu:' dropdown, there is a label 'h: nspecifikováno' and a dropdown arrow. At the bottom of the form are two buttons: 'Hledat' and 'Smazat formulář'.



# Ne každý je lemmatizovaný...

Korpus: intercorp\_ar

Typ dotazu: Základní

Dotaz: Fráze

Subkorpus

- Základní
- Fráze
- Slovní tvar
- Podřetězec
- CQL

# Přidávání paralelních korpusů

Korpus:

Typ dotazu:

Dotaz:

Zarovnané korpusy

korpus

- intercorp\_ar
- intercorp\_be
- intercorp\_bg
- intercorp\_ca
- intercorp\_cs
- intercorp\_da
- intercorp\_el
- intercorp\_en**
- intercorp\_es
- intercorp\_et
- intercorp\_fi
- intercorp\_fr
- intercorp\_hi
- intercorp\_hr
- intercorp\_hu
- intercorp\_it
- intercorp\_lt
- intercorp\_lv
- intercorp\_mk
- intercorp\_mt



# Přidávání dalších paralelních korpusů

Korpus:

Typ dotazu:

Dotaz:

Zarovnané korpusy

korpus

Typ dotazu:

Dotaz:


zobrazit prázdné řádky



# Paralelní korpus – zadání dotazu

Korpus:


Typ dotazu:

Podřetězec:  

Zarovnané korpusy

korpus

Typ dotazu:

Podřetězec:  

zobrazit prázdné řádky



# Paralelní – výsledky typ dotazu: podřetězec

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

strana  ze 65  [další](#) | [poslední](#)

	<a href="#">intercorp_de</a>	<a href="#">intercorp_en</a>
<a href="#">Per Anhalter durch die Galaxis</a>	<p> Zaphod <b>liebte</b> die große Show : darin war er der Größte . </p>	<p> Zaphod <b>loved</b> effect : it was what he was best at . </p>
<a href="#">Per Anhalter durch die Galaxis</a>	<p> Das passte den Dentrassis ausgezeichnet in den Kram , denn sie <b>liebten</b> das vagonische Geld , das eine der härtesten Währungen im Weltall ist , aber die Vogonen konnten sie nicht leiden .	<p> This suited the Dentrassis fine , because they <b>loved</b> Vagon money , which is one of the hardest currencies in space , but loathed the Vogons themselves .
<a href="#">Per Anhalter durch die Galaxis</a>	Zu Hause war diese Industrie auf dem Planeten Magrathea , wo Hyperraum-Ingenieure durch weiße Löcher im All Materie ansaugten und sie in <b>liebevoll</b> gestaltete Traumplaneten verwandelten - Goldplaneten , Platinplaneten , Weichgummiplaneten mit Massen von Erdbeben - , die selbst den verwöhntesten Ansprüchen der reichsten Männer der Galaxis genügten . </p>	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 <b>lovingly</b> made to meet the exacting standards that the Galaxy 's richest men naturally came to expect . </p>
<a href="#">Die Brücke über die Drina</a>	Das verdirbt das Spiel und ruft Enttäuschung und Unmut bei denen hervor , die das Spiel der Phantasie <b>lieben</b> , die Ironie hassen und glauben , mit geduldigem Hinschauen könne man wirklich etwas sehen und erleben . </p>	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 <b>loved</b> the play of imagination , hated irony and believed that by looking intently they could actually see and feel something .
<a href="#">Die Brücke über die Drina</a>	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 <b>liebten</b> , 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 . </p>	When the next day , in the cloudy dawn , they looked down from the hillside on the town that they <b>loved</b> 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 . </p>

# Volby zobrazení

Po vyhledání konkordance můžeme upravovat možnosti zobrazení – lze si nastavit např. zobrazování jména autora či tagu u KWICu.

(KWIC – červeně zvýrazněné klíčové slovo, key word in context)

Korpus: intercorp\_en Popis: Angličtina - korpus intercorp\_en, verze 6 z 8. 4. 2013

Možn. zobrazení

Atributy	Struktury	Reference
<input checked="" type="checkbox"/> word	<doc>	doc.wordcount
<input type="checkbox"/> lemma	<div>	div.author
<input checked="" type="checkbox"/> tag	<p>	div.title
<input type="checkbox"/> lc	<s>	div.publisher
<input type="checkbox"/> lemma_lc		div.pubplace
Zobrazit atributy		
<input type="radio"/> u všech tokenů		
<input checked="" type="radio"/> pouze u KWIC		

# 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 **označkových** (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
<b>základní</b>	bez	více	zadáme-li tvar lemmatu, vyhledá celé paradigma
<b>lemma</b>	s	jen 1	lze specifikovat sl. druh ( <i>stát</i> jako sloveso)
<b>fráze</b>	s	více	konkrétní slovní tvary
<b>slovní tvar</b>	s	jen 1	lze specifikovat sl. druh ( <i>pří</i> jako podst. jm. – od <i>pře</i> )



# Typy dotazů – dokončení

typ dotazu	s / bez RE (regulárních výrazů)	počet slov	další
<b>podřetězec</b>	s	1 řetězec	vyhledá např. <i>mrsk</i> – všechny odvozeniny slov <i>mrskat</i> , <i>mrsknout</i> , <i>smrsknout</i> , <i>Zámorsk</i> , i překlady typu <i>mrskem</i>
<b>CQL</b>	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

Korpus:

Typ dotazu:

Slovní tvar:

- ▼ Synchronní psané korpusy
  - ▼ řada SYN
    - [syn](#)
    - [syn2010](#)
    - [syn2009pub](#)
    - [syn2006pub](#)
    - [syn2005](#)
    - [syn2005wsbr](#)
    - [syn2000](#)
  - ▶ specializované
  - ▶ ke slovníkům
- ▼ Synchronní mluvené korpusy
  - ▼ řada ORAL
    - [oral2013](#)
    - [oral2008](#)
    - [oral2006](#)
  - ▶ specializované
- ▶ Diachronní korpusy
- ▶ Cizojazyčné korpusy
- ▶ Cizojazyčné korpusy webové
- ▶ Paralelní korpus InterCorp



# 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



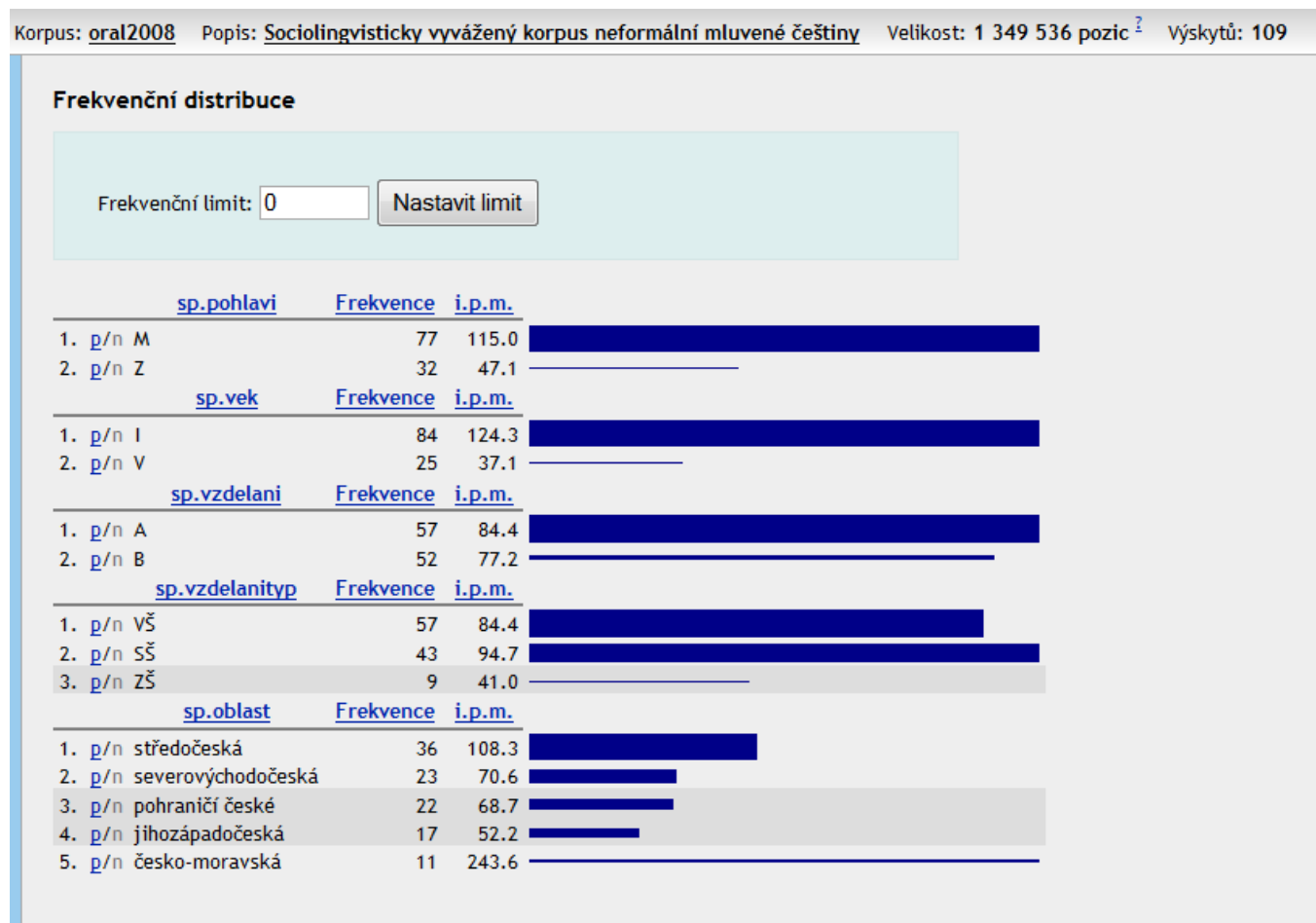




# STATISTIKY / Frekvenční distribuce



# Frekvenční distribuce – ORAL2008



# 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 %
<b>celkem</b>	<b>257 868</b>	<b>100 %</b>

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

lemma	absolutní frekvence	podíl
he	95 875	65,8%
she	52 148	34,2 %
<b>celkem</b>	<b>148 023</b>	<b>100 %</b>

# 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
<i>his</i> (DPS, PNP, UNC) (PP\$)	BNC InterCorp	409 825 117 914	<b>3 684</b> <b>1 826</b>
<b>originál: <i>his</i></b>	<b>InterCorp EN</b>	<b>54 685</b>	<b>7 774</b>
<i>her</i> (DPC, UNC, VVG) (PP\$, PP)	BNC InterCorp	23 376 77 410	<b>1 828</b> <b>1 199</b>
<i>jeho</i>	SYN	2 601 136	<b>1 658</b>
<i>její</i>	SYN	1 163 534	<b>742</b>

# Pozor na skladbu korpusů

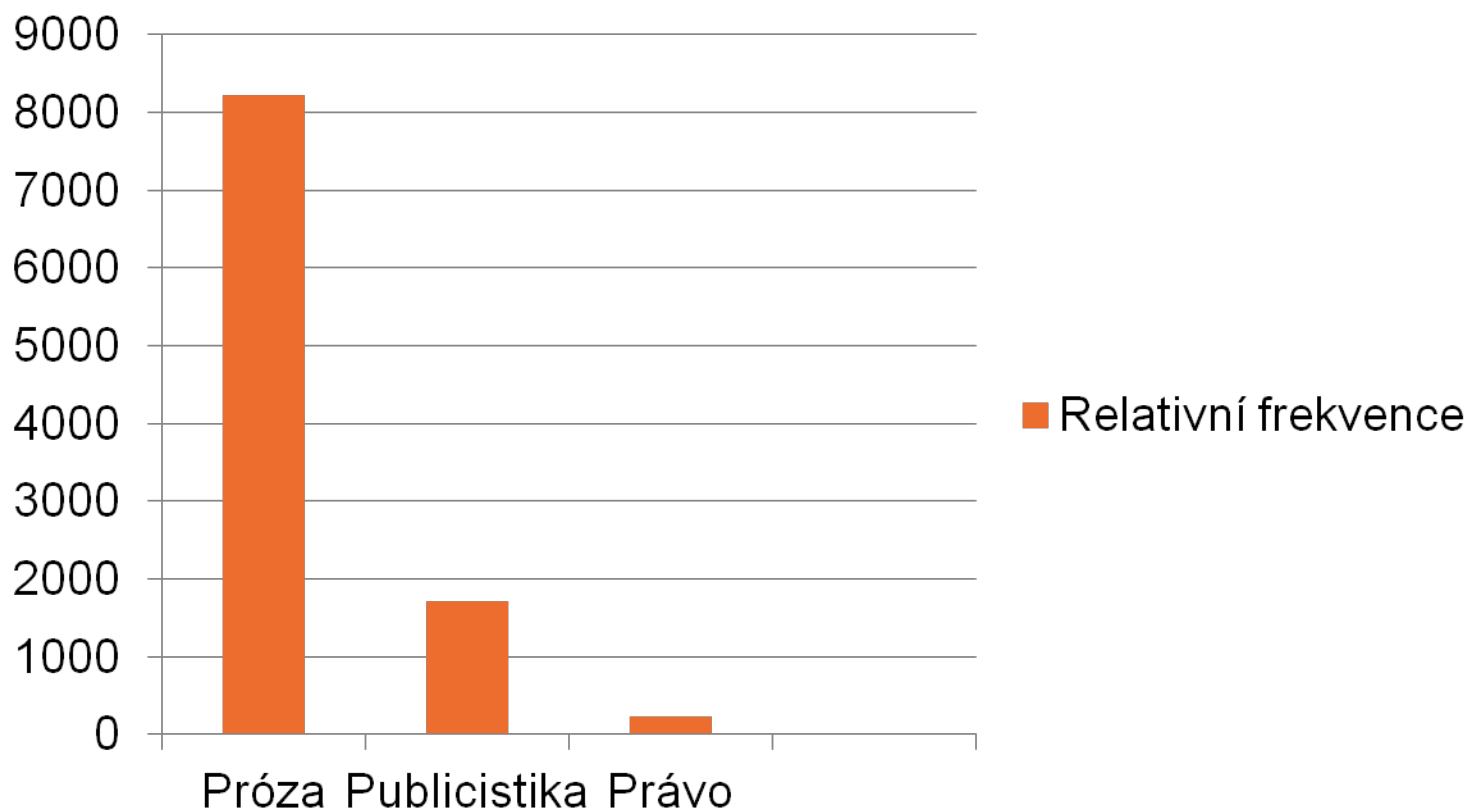
<u>Text type</u>	<u>Frekvence</u>	<u>i.p.m.</u>
1. p/n Written books and periodicals	375 868	4185.4
2. p/n Written miscellaneous	13 916	1685.8
3. p/n Spoken context-governed	7 304	1066.3
4. p/n Spoken demographic	6 762	1382.6
5. p/n Written-to-be-spoken	5 975	4137.4
<u>Publication date</u>	<u>Frekvence</u>	<u>i.p.m.</u>
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
<u>Domain for written corpus texts</u>	<u>Frekvence</u>	<u>i.p.m.</u>
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

Výskyty zájmeny *his* v InterCorpu (anglické složce) v 1 milionu slov





# 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](#)

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

	<a href="#">intercorp_en</a>		<a href="#">intercorp_cs</a>
<a href="#">brown-chut_lasky,ANO,en</a>	Have you got balls enough to face me like a man ? Or are we gonna continue this silly game of hide-and-peek ? " </p><p> Following a short silence , a voice came to him from the other side of the wall . " Threadgill ? " </p>	<a href="#">brown-chut_lasky,NE,en</a>	Máš dost odvahy , aby ses mi postavil jako chlap ?
<a href="#">styblova-skalpel_pros,NE,cs</a>	<p> How had it got here ? What ? Mr Materna gave it to us ?	<a href="#">styblova-skalpel_pros,ANO,cs</a>	<p> Jak se to tu octlo ?
<a href="#">munro-utek,ANO,en</a>	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 ? </p>	<a href="#">munro-utek,NE</a>	Cožpak věděla s jistotou , že si ji nepořídili jako náhradu ?
<a href="#">Brown-zdravim_temnoto,ANO,en</a>	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 ! " </p><p> Gavin sidestepped , blocking her view of the others and forcing her to look at him . " Is Janey really missing ? " </p>	<a href="#">Brown-zdravim_temnoto,NE,en</a>	<p> Gavinovi přes rameno uviděla hlouček nějakých známých , zamávala na ně a zavolala : „ Ahoj , zrovna jsem přijela z Francie , počkejte , až vám budu vyprávět ! “ </p>
<a href="#">woolfova-dallowayova,ANO,en</a>	<p> What had she got in her work-box ?	<a href="#">woolfova-dallowayova,NE,en</a>	<p> Co ještě má v košíčku na šití ?
<a href="#">kohout-snezim,NE,cs</a>	I 'll lose , of course , but first I have to get some guarantee from her that will reassure me . </p><p> " 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 ! </p><p> " What kind of guy is he ? "	<a href="#">kohout-snezim,ANO,cs</a>	( Prohraju , ovšemže , ale předtím z ní musím dostat záruky , které mě aspoň trochu uklidní . ) </p>
<a href="#">adams-stoparuv_pruvodc,ANO,en</a>	" 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 ? " </p><p> 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	<a href="#">adams-stoparuv_pruvodc,NE,en</a>	" 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 ? " </p>





# PODMÍNKY A SUBKORPUSY



# Podmínky: v rámci jedné věty

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

REF: 65 | Výsledek je promíchán

<code>&lt;/p&gt;&lt;p&gt;</code> What	've you got there ?	<code>&lt;/p&gt;&lt;p&gt;</code> VÉNA har
said Thorin , "	have n't you got a map ?	and did n't you he;
ack . <code>&lt;/p&gt;&lt;p&gt;</code> "	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
ne , Mr Dixon ;	have you got a minute to spare ?	' <code>&lt;/p&gt;&lt;p&gt;</code> 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 ?	" <code>&lt;/p&gt;&lt;p&gt;</code> " You '
<code>/p&gt;&lt;p&gt;</code> " When	have you got to go ?	" he asked in an
> LUCY : What	have you got to do ?	<code>&lt;/p&gt;&lt;p&gt;</code> ( She bur
><p> " Lozada !	Have you got balls enough to face me like a man ?	Or are we gonna c
<code>/p&gt;&lt;p&gt;</code> " What	have you got there ?	" she asked curio
it ; exactly how	had she got him to come inside her room ?	And did n't she ha
<code>&lt;/p&gt;&lt;p&gt;</code> ' What	have you got in the basket , Pippi ?	' asked Annika . '
iment 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 ?	' <code>&lt;/p&gt;&lt;p&gt;</code> Dixon sv
I 'm back now .	Have you got a job for me ?	" <code>&lt;/p&gt;&lt;p&gt;</code> Armans
n countries , or	have I got the wrong person ?	<code>&lt;/p&gt;&lt;p&gt;</code> Now , Eur
kson ? <code>&lt;/p&gt;&lt;p&gt;</code>	Have you got a pet animal ?	<code>&lt;/p&gt;&lt;p&gt;</code> Will you §
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> [tag!="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 <div author="Milne.*" />`
  - 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 <div author="Woolf.*" />`
- pozor: i.p.m. (80 výskytů) vztaženo k celému korpusu!



# Vytváření subkorpusů

- ***Subkorpus – Vytvořit nový – Vlastní within podmínka***
  - within `<div author="Woolf.*" />`
    - 186 222 tokenů
  - Hledat v: *Dostupné subkorpora*
    - 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="l.*"]within <opus srclang="" />`
  - můžeme porovnat např. s citoslovci v jazyce překladu:
    - `[tag="l.*"]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](mailto:olga.richterova@ff.cuni.cz)  
[michal.kren@ff.cuni.cz](mailto:michal.kren@ff.cuni.cz)

