



ČESKÝ NÁRODNÍ
KORPUS

Bychom, nebo *bysme*?

Jak snadno a rychle porovnat varianty v psané a mluvené češtině
a naučit studenty je rozlišovat

23. 3. 2019
Lucie Lukešová



CO NÁS DNES ČEKÁ

1. Něco málo o jazykových korpusech
 - co to je a jak se to používá
2. Korpusový portál a volně dostupné nástroje
 - webový portál www.korpus.cz
3. Korpusový nástroj SyD
 - porovnání dvou a více slov z hlediska jejich užití
 - volba vhodného výrazu pro danou komunikační situaci
 - srovnání mluveného a psaného jazyka
 - výběr častější pravopisné varianty
4. Portál Pro školy s pracovními listy

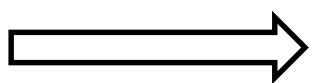


CO JE TO VŮBEC TEN KORPUS...?



Jazykový korpus...

- je rozsáhlý soubor nejrůznějších textů (psaných i původně mluvených) v elektronické podobě
- můžeme prohledávat online (hledat slova, fráze, části slov apod.)
- zahrnuje další informace o textech i o jednotlivých slovech
- zobrazuje slova v jejich přirozeném kontextu
- ukazuje, jak skutečně píšeme a mluvíme



pozor, korpus nejsou *Pravidla českého pravopisu*!



Jak korpus vzniká?

autoři textů



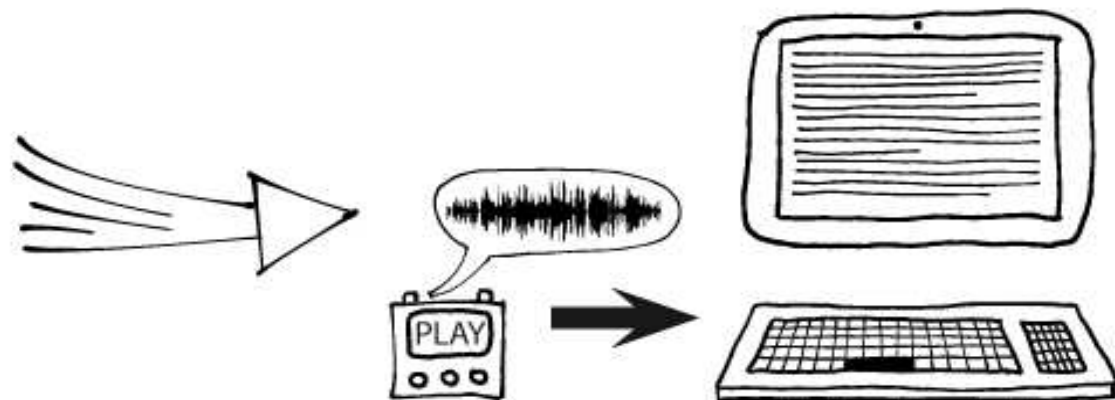
nakladatelé a vydavatelé



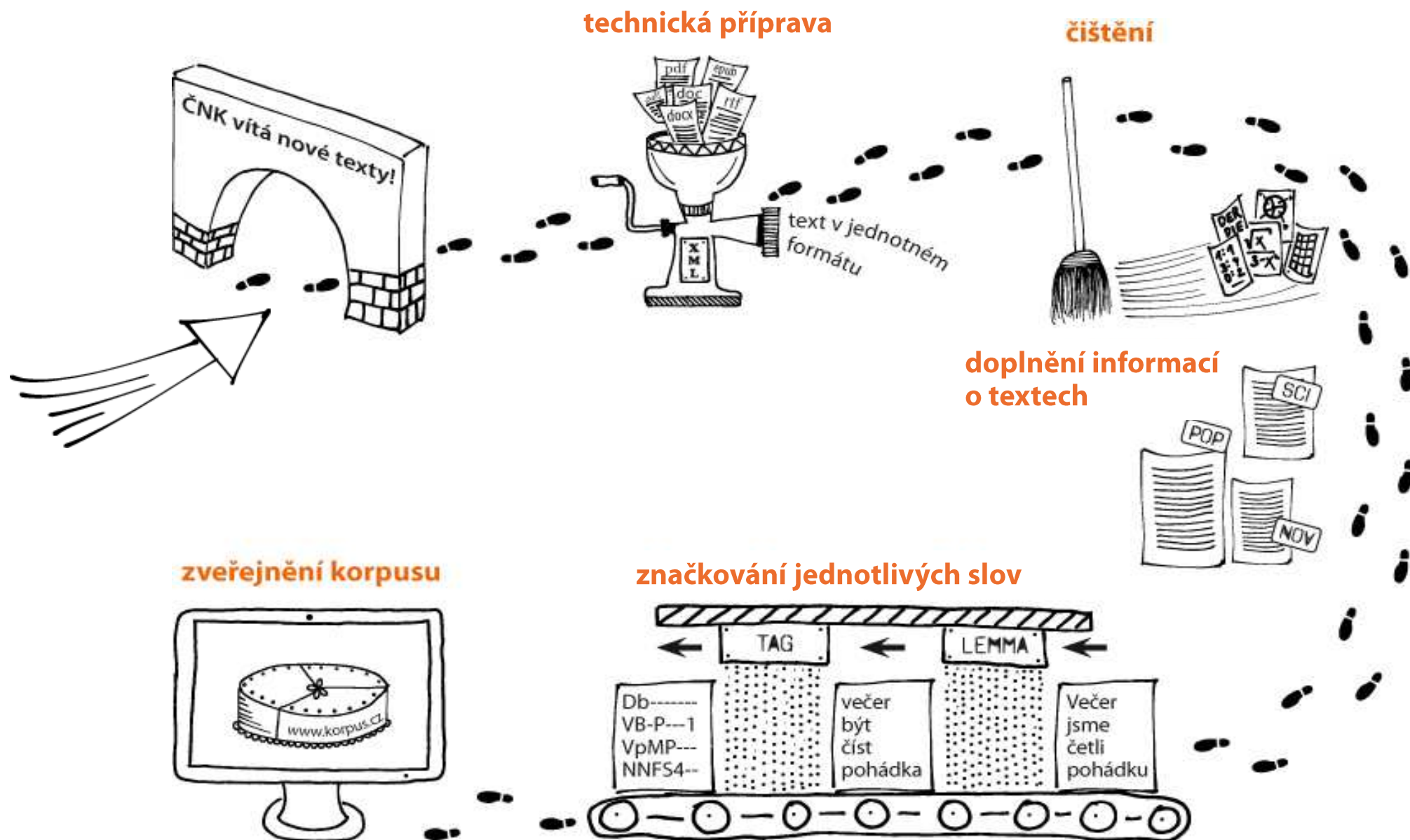
sběr mluveného jazyka



přepis nahrávek



Jak se z těchto textů stane korpus?



Jak takový „nahatý“ korpus vypadá?

```
<opus autor="Doyle, Arthur Conan" nazev="Příběhy Sherlocka Holmese" nakladatel="Mladá fronta" mistovyd="Praha"
rokvyd="1971" isbnissn="" preklad="Henzl, V. - Zábrana, J. - Wolfová, Z." srclang="ENG" txttype_group="beletrie"
txttype="NOV" genre="CRM" med="B" id="pribshho">
<doc id="1">
...
<s id="10">
Když      když      J,-----
školení školení NNNS4-----A-----
skončilo      skončit VpNS---3R-AA---P
,            ,      Z:-----
přidělili      přidělit      VpMP---3R-AA---P
mne      já      PP-S4--1-----
k      k      RR--3-----
Pátému Pátý      NNMS3-----A-----
northumberlandskému      northumberlandský      AAIS3-----1A-----
střeleckému      střelecký      AAIS3-----1A-----
pluku      pluk      NNIS3-----A-----
jako      jako      J,-----
pomocného      pomocný AAMS4-----1A-----
chirurga      chirurg NNMS4-----A-----
.            .      Z:-----
</s>
...
</doc>
...
</opus>
<opus>
...
</opus>
```

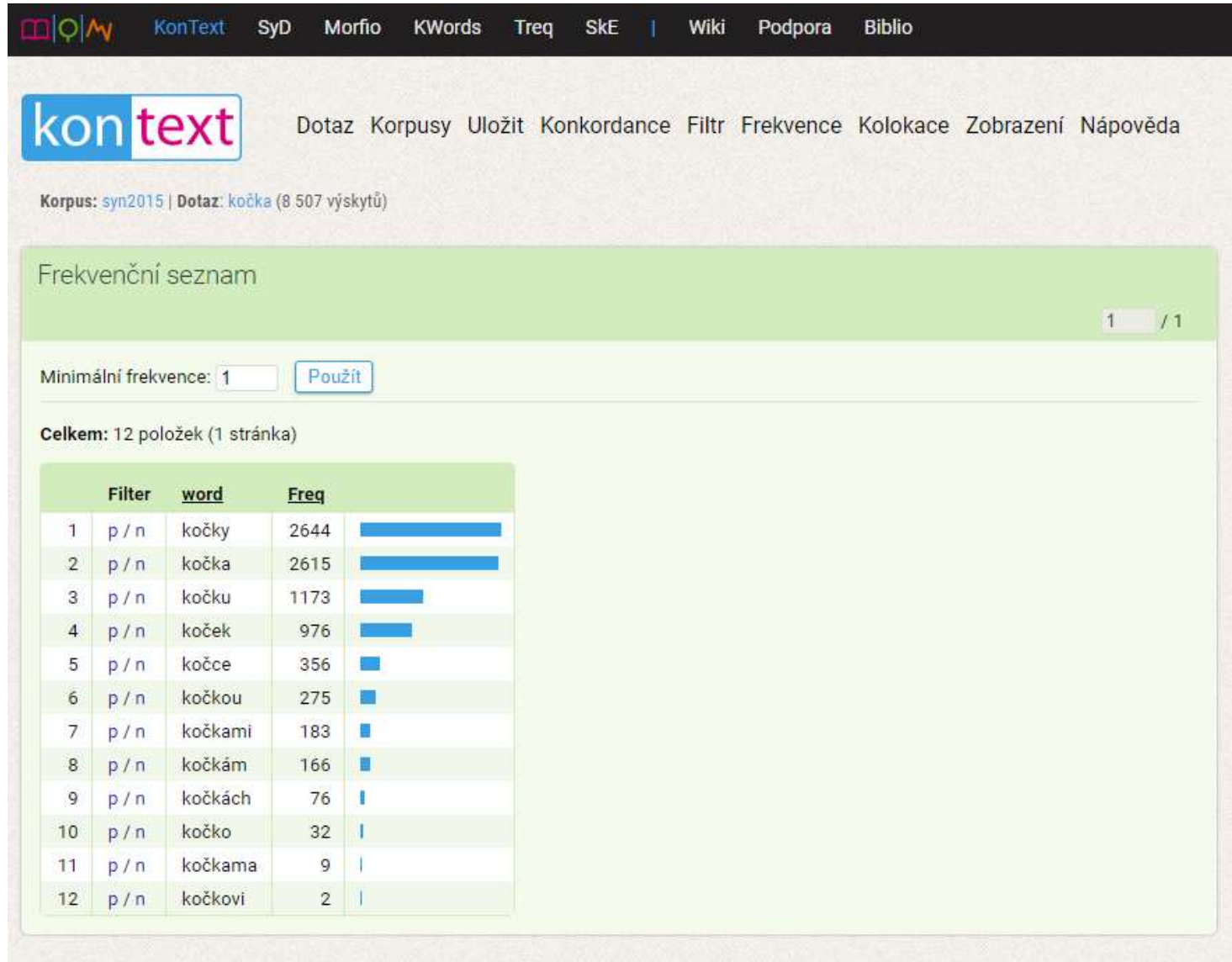
A jak ho pak používáme my?

1. prohledáváme jej přímo v **korpusovém vyhledávači** (např. KonText)

The screenshot shows the KonText web interface. At the top, there's a navigation bar with links like SyD, Morfio, KWords, Treq, SkE, Wiki, Podpora, and Biblio. The main header includes the KonText logo and a search bar. Below the search bar, it shows the current search: 'Korpus: syn2015 | Dotaz: kočka (8 507 výskytů)'. The results section displays a list of 20 items, each with a checkbox, a source label (e.g., NOV: próza, POP: populárně naučná literatura), a snippet of text, and the word 'kočka' with its grammatical information (e.g., /kočka/NNFS1-----A-----). The interface also includes filters for 'Výběr řádků' and 'Atributy', and a pagination bar at the bottom showing '1 / 426'.

A jak ho pak používáme my?

1. prohledáváme jej přímo v **korpusovém vyhledávači** (např. KonText)



The screenshot shows the KonText web interface. At the top, there is a navigation bar with links: KonText, SyD, Morfio, KWords, Treq, SkE, Wiki, Podpora, and Biblio. Below this, the KonText logo is displayed, followed by a menu: Dotaz, Korpusy, Uložit, Konkordance, Filtr, Frekvence, Kolokace, Zobrazení, and nápověda. The main content area shows the search results for the word 'kočka' in the 'syn2015' corpus, with 8,507 occurrences. A section titled 'Frekvenční seznam' (Frequency list) is highlighted in green. It includes a filter for 'Minimální frekvence' (Minimum frequency) set to 1, and a 'Použít' (Use) button. Below this, it states 'Celkem: 12 položek (1 stránka)' (Total: 12 items (1 page)). A table lists the top 12 most frequent forms of the word, sorted by frequency. Each row includes a rank, a filter (p/n), the word form, the frequency, and a horizontal bar chart representing the frequency.

	Filter	word	Freq	
1	p / n	kočky	2644	
2	p / n	kočka	2615	
3	p / n	kočku	1173	
4	p / n	koček	976	
5	p / n	kočce	356	
6	p / n	kočkou	275	
7	p / n	kočkami	183	
8	p / n	kočkám	166	
9	p / n	kočkách	76	
10	p / n	kočko	32	
11	p / n	kočkama	9	
12	p / n	kočkoví	2	

seznam tvarů
od nejčastějšího

A jak ho pak používáme my?

1. prohledáváme jej přímo v **korpusovém vyhledávači** (např. KonText)

seznam typů textů, v
nichž se píše o **kočce**

Frekvenční seznam

Minimální frekvence:

Celkem: 3 položky

	Filter	doc.txttype_group	Freq	i.p.m.	
1	p / n	NFC: oborová literatura	3526	89.53	
2	p / n	FIC: beletrie	3315	79.65	
3	p / n	NMG: publicistika	1666	41.92	

Celkem: 12 položek

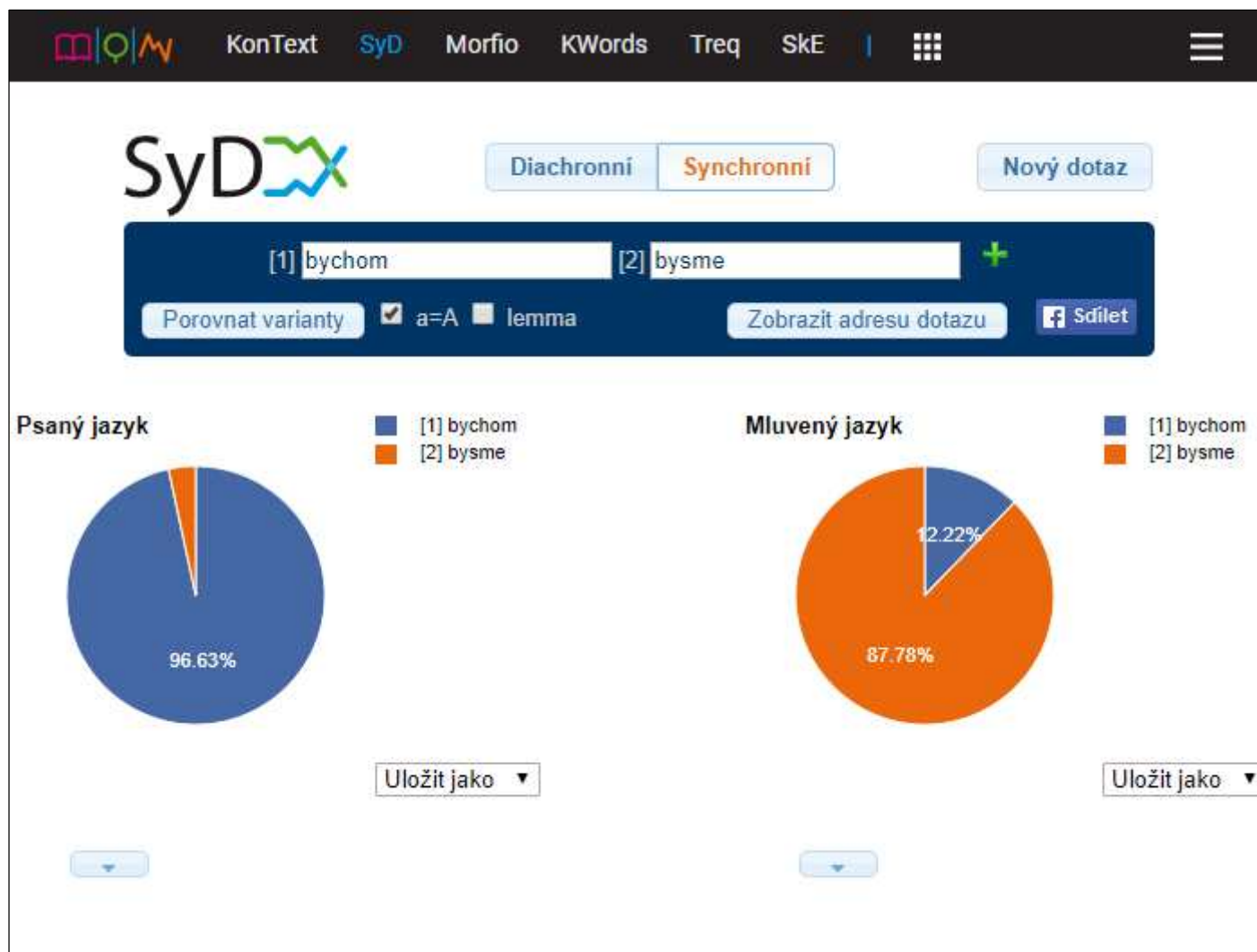
	Filter	doc.txttype	Freq	i.p.m.	
1	p / n	POP: populárně naučná literatura	3194	242.42	
2	p / n	NOV: próza	2549	78.13	
3	p / n	LEI: volnočasová publicistika	1184	73.83	
4	p / n	NEW: tradiční publicistika	482	20.33	
5	p / n	COL: kratší próza	458	73.15	
6	p / n	VER: poezie	222	201.03	
7	p / n	SCI: odborná literatura	136	10.51	
8	p / n	MEM: memoáry, autobiografie	135	28.22	
9	p / n	SCR: drama	68	52.25	
10	p / n	PRO: profesní literatura	57	7.05	
11	p / n	X: neuvedeno	18	54.75	
12	p / n	ADM: administrativa	4	9.95	

Celkem: 8 položek

	Filter	doc.genre_group	Freq	i.p.m.	
1	p / n	X: neuvedeno	4981	61.22	
2	p / n	NAT: přírodní vědy	2947	349.93	
3	p / n	SSC: sociální vědy	165	20.17	
4	p / n	MEM: memoáry, autobiografie	135	28.22	

A jak ho pak používáme my?

2. využíváme korpusová data v **korpusových nástrojích** (např. SyD)

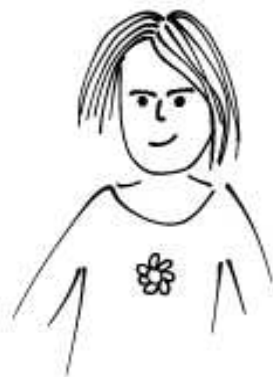


Kdo korpus používá a na co?



jazykovědci

- píší **slovníky**
- tvoří **mluvnice**
- zkoumají **proměny jazyka**
- porovnávají **různé jazyky**



studenti

- učí se, jak vypadá **současný jazyk**
- porovnávají **užití slov**
- analyzují **texty**
- píší **práce**



učitelé

- tvoří **cvičení**
- ukazují studentům, co je **v jazyce obvyklé** a co ne
- hledají aktuální **informace o češtině**



překladaelé



a další...

- hledají **typické fráze** v různých jazycích
- vyhledávají, **jak se něco** obvykle **překládá**

Jaké korpusy pro češtinu máme a co v nich je?



psané: noviny & časopisy, odborné knihy i romány (řada **SYN**); historická čeština (**Diakorp**)



mluvené: přepisy + nahrávky skutečných rozhovorů, získaných po celé ČR (řada **ORAL, ORTOFON**)



paralelní: originální texty a jejich překlady v češtině a 39 dalších jazycích (**InterCorp**)



specializované: KSK, Totalita, Čapek, Hrabal...



Seznam všech korpusů na www.korpus.cz

KonTextSyDMorfioKWordsTregSkE**Wiki**FodporaBiblio

Lucie LukešováOdhlášení | English

Přihlásit se

WIKI
ČESKÝ NÁRODNÍ
KORPUS

Navigace

- Manuál
- Korpusy ČNK**
- Kurz práce s ČNK
- Slovníček pojmů
- Seznamy

Tisk/export

Verze k tisku

Nástroje

Stránky odkazující sem

Poslední úpravy

Správa médií

Hlavní stránka

Trvalý odkaz

Citovat tuto stránku

Jazyky

- Česky (Česky)
- English (English)

Licence

ČístZdrojový kód stránkyStarší verze

Jaké korpusy zpřístupňuje Český národní korpus?

Projekt [Český národní korpus](#) zpřístupňuje řadu různých [korpusů](#), jejich okomentovaný přehled najdete v sekci týkající se jeho [složení](#). Celkový rozsah textů, který ČNK v korpusech zpracoval a zpřístupnil, přesahuje 3,6 miliardy slov v českých jednojazyčných a 1,5 miliardy slov v cizojazyčných [paralelních korpusech](#).

Korpusy slouží jak laikům, tak odborníkům jako východisko pro [různé druhy](#) empirického poznávání jazyka. Pravidla jejich užívání se řídí jednotnými zásadami, které uživatelé schvalují při bezplatné [registraci](#), a podléhají běžným zvyklostem pro [citování](#) vědeckých pramenů.

Přehled dostupných korpusů

Korpusy psaného jazyka (synchronní)

korpus	velikost (počet slov)	lemmatizace	morfologické značky	rok zveřejnění ¹⁾	charakteristika korpusu
Obecné korpusy					
SYN (verze 7)	4,255 mld.	✓	✓	2010	verzovaný korpus, spojující synchronní psané korpusy řady SYN a další, dosud nezveřejněné texty
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
Specializované korpusy					
CZESL-PLAIN	2 mil.	X	X	2012	žakovský korpus češtiny nerodilých mluvčích
CZESL-SGT	960 tis.	✓	✓	2014	žakovský korpus češtiny nerodilých mluvčích s metadaty a automatickou anotací
FicTree	135 tis.	✓	✓	2017	manuálně syntakticky anotovaný korpus beletrie
FSC2000	100 mil.	✓	X	2004	upravený SYN2000, referenční zdroj Frekvenčního slovníku češtiny
JEROME	85 mil.	✓	✓	2013	jednojazyčný srovnatelný korpus pro výzkum překladové češtiny
Koditex	10,8 mil.	✓	✓	2018	korpus vytvořený pro účely multidimenzionální analýzy češtiny
KH-DOPISY	500 tis.	X	X	2017	Korpus korespondence Karla Havlíčka
KSK-DOPISY	800 tis.	X	X	2006	Korpus soukromé korespondence: přepisy ručně psaných dopisů z let 1990–2004
LINK	1,8 mil.	✓	✓	2010	korpus sestavený z odborných lingvistických textů
ORWELL	80 tis.	✓	✓	2003	ručně označovaný korpus Orwellova románu 1984 🐼
SKRIPT2012	590 tis.	✓	✓	2013	korpus školních písemných prací
Korpusy mluveného jazyka (synchronní)					
korpus	velikost (počet slov)	lemmatizace	morfologické značky	rok zveřejnění	charakteristika korpusu
Obecné korpusy					



JAK S KORPUSY PRACOVAT?



www.korpus.cz

všechny naše **nástroje** najdete nahoře na černé liště

máme i svou **Wikipedii** s informacemi



NA CO SE HODÍ KTERÝ NÁSTROJ?

- **KonText:** chci vidět **slovo použité v kontextu**, zjistit **informace o jeho použití** nebo si vytvořit **seznam nejčastějších slov** atd. atd.
- **SyD:** chci si **porovnat dvě nebo více slov**, jestli se používají stejně, které je častější, kdy se hodí použít, kdo je používá apod.
- **KWords:** chci zjistit, jaká jsou **typická (klíčová) slova v mém textu** v porovnání s ostatními texty/korpusem
- **Treq:** chci najít **překlad slova nebo fráze** v češtině/angličtině a dalších 39 jazycích na základě přeložených textů



NA CO SE HODÍ KTERÝ NÁSTROJ?

- **KonText:** chci vidět slovo použité v kontextu, zjistit informace o jeho použití nebo si vytvořit seznam nejčastějších slov atd. atd.
- **SyD:** chci si **porovnat dvě nebo více slov**, jestli se používají stejně, které je častější, kdy se hodí použít, kdo je používá apod.
- **KWords:** chci zjistit, jaká jsou typická (klíčová) slova v mém textu v porovnání s ostatními texty/korpusem
- **Treq:** chci najít překlad slova nebo fráze v češtině/angličtině a dalších 39 jazycích





SyD – porovnáváme slova

O SyDu

- Co SyD umí?
 - rychle porovnat vzájemně si konkurující jazykové jevy (tj. pravopisné, morfologické, lexikální či stylistické varianty – slova i fráze)
- Jak ho můžeme využít ve výuce?
 - na zjištění rozdílů mezi psanou formální a spontánní mluvenou češtinou
 - na ověření frekventovanosti různých výrazů (*filozofie*, nebo *filosofie*?)
 - na vizualizaci regionálních rozdílů v mluveném jazyce apod.



syd.korpus.cz

Diachronní

Synchronní

ÚČNĚK



Korpusový průzkum variant

[1]

[2]

+

☒ a=A ☐ lemma



SyD je určen pro všestranný průzkum variant jak v současném jazyce - **synchronní část**, tak v průběhu jeho vývoje - **diachronní část**. K analýze využívá data mluvených a psaných korpusů projektu Českého národního korpusu. Stačí zadat dvě varianty (nebo víc) jednoho jevu, které si vzájemně konkurují (např. *ačkoliv* × *ačkoliv*, *už* × *již*, *mimoto* × *mimo to*, *sama sebe* × *sebe sama*, *stále* × *pořád* × *furt* apod.).

Vyhodnocení složitějších dotazů, jejichž podoba může využít všech možností dotazovacího jazyka CQL, může trvat i několik desítek sekund.

Poslední položené dotazy:

takhl.* × takl.*
Vánoc.* × vánoc.*
blbost × blbina (lemma)
pepa × josef
ohánět se

Synchronní část: současná čeština



Diachronní část: historická čeština

KonText Park SyD Morfio KWords Wiki Podpora Biblio

SyD

Diachronní

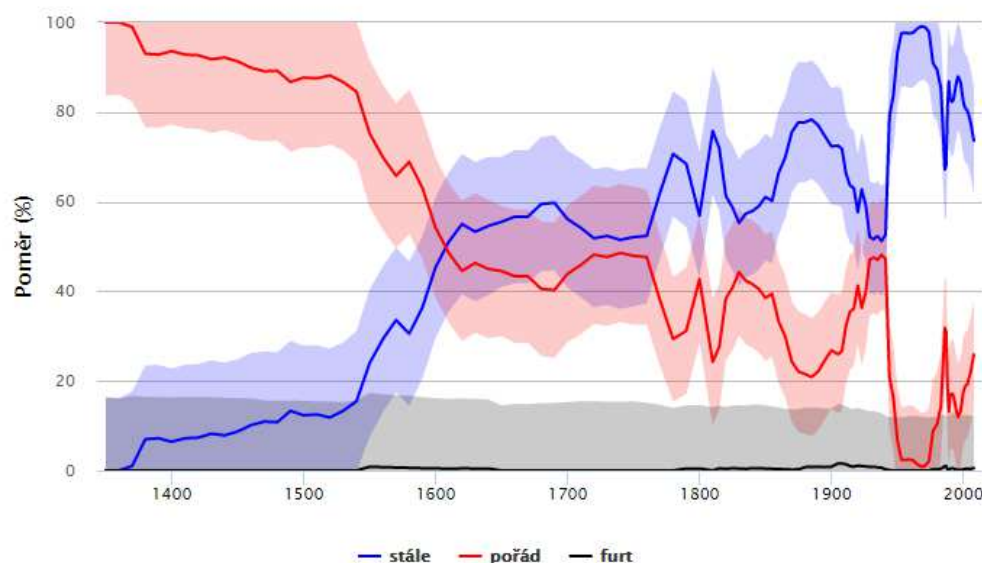
Synchronní

Nový dotaz

[1] stále [2] pořád [3] furt

Okno: proměnlivé Krok: proměnlivé

Porovnat varianty ☒ a=A ☐ lemma Zobrazit adresu dotazu



Klouzavý průměr: 1 ☒ zobrazit chybu Uložit jako

Celá historie

14.-18. stol.

19. stol.

1900-1989

1990-2009

Zkoumání proměn poměru variant v čase je založeno na korpusu *Diakon*, který je sestaven na základě textů korpusu SYN a diachronní složky ČNK tak, aby v rámci současných možností co nejlépe pokrýval celé období existence psané češtiny. Uživatelé je třeba upozornit na skutečnost, že texty diachronní složky ČNK zařazené do korpusu *Diakon* z velké části nebyly dosud zkorigovány, a výsledky vyhledávání konkurence variant nelze proto považovat za zcela přesné, i když už v současné době umožňují získat vysoce spolehlivou představu o celkovém vývoji. *Diakon* zahrnuje texty od nejstarších památek ze 14. století (nejstarším je *Alexandreida*, datovaná do roku 1310) až po publicistiku z let 1991-2009. Vzhledem k proměnám písemné produkce, kterými čeština za 7 století své existence prošla, není možné klást na takový korpus stejné požadavky stylové a žánrové vyváženosti, jaké máme na korpusy češtiny současné.

Úvodní graf prezentuje tendence vývoje poměru variant v průběhu času. Každá linie reprezentující jednu variantu je obklopena stejnobarevnou oblastí, jejíž šíře značí spolehlivost naměřených dat (čím širší oblast, tím větší možná chybovost výsledku).

Jak se liší použití zájmen *tenhle* a *tento*?

Píšeme častěji *filosofie*, nebo *filozofie*?

S čím se kombinují synonyma *statečný* a *odvážný*?

Jak se liší použití slov *kluk* a *chlapec* nebo *holka* a *dívka*?

Kde v ČR se říká *vokno* a kde ne?



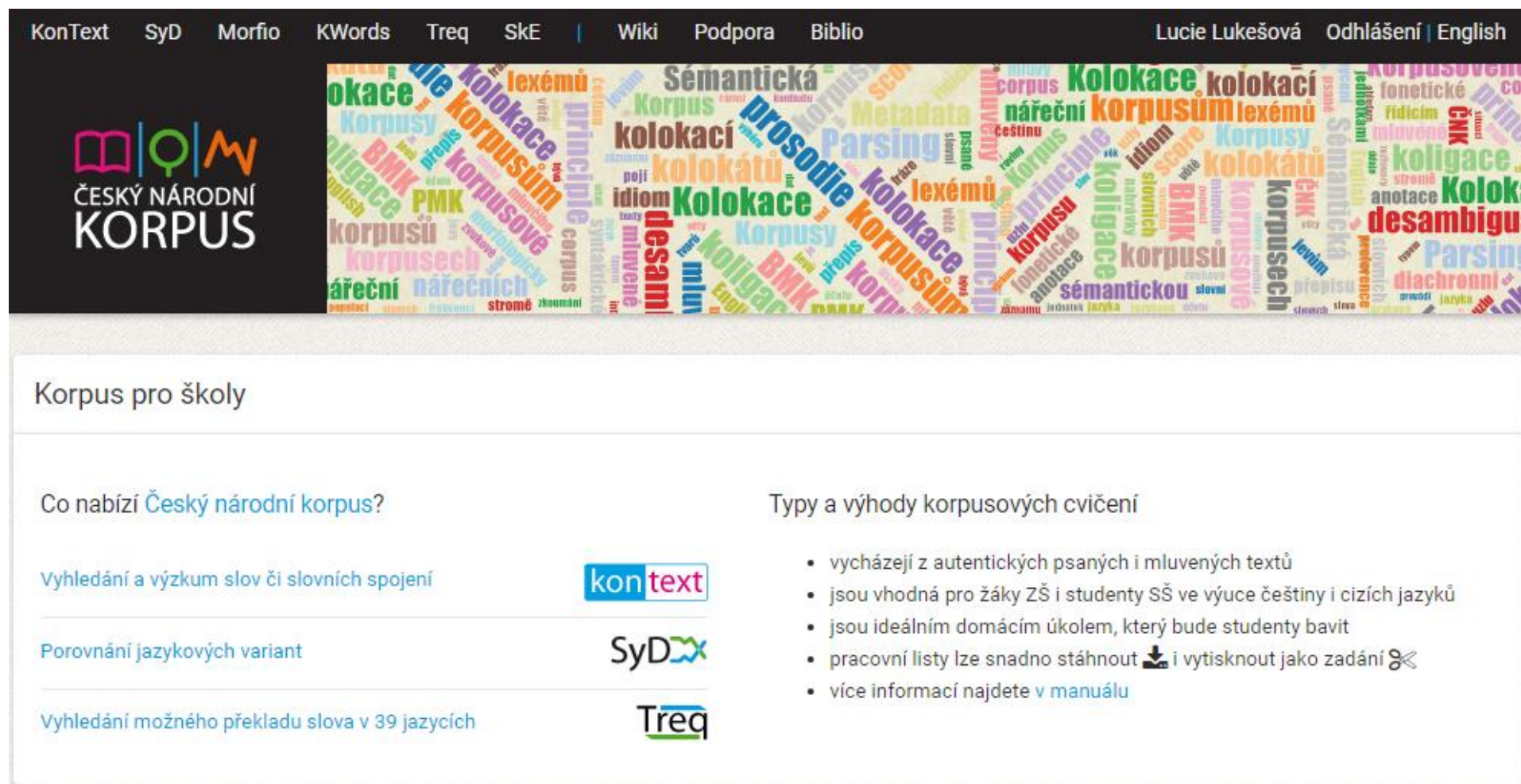


Portál Pro školy



www.korpus.cz/proskoly

- repositář cvičení a nápadů pro využití korpusů ve výuce



KonText SyD Morfio KWords Treq SkE | Wiki Podpora Biblio Lucie Lukešová Odhlášení English

ČESKÝ NÁRODNÍ KORPUS

Korpus pro školy

Co nabízí Český národní korpus?

- Vyhledání a výzkum slov či slovních spojení 
- Porovnání jazykových variant 
- Vyhledání možného překladu slova v 39 jazycích 

Typy a výhody korpusových cvičení

- vycházejí z autentických psaných i mluvených textů
- jsou vhodná pro žáky ZŠ i studenty SŠ ve výuce češtiny i cizích jazyků
- jsou ideálním domácím úkolem, který bude studenty bavit
- pracovní listy lze snadno stáhnout  i vytisknout jako zadání 
- více informací najdete [v manuálu](#)

www.korpus.cz/proskoly

- vhodná cvičení lze hledat podle štítků

Hledejte pomocí štítků
Kliknutím vyberte, dalším kliknutím zrušte

ČJ ČCJ AJ NJ FJ papír počítač projektor slovo tvorba tvarosloví slovní zásoba skladba pravopis stylistika mix
A1 A2 B1 B2 C1 rodilý 6.-7. roč. 8.-9. roč. SŠ pozorování porovnávání doplňování spojování hledání mix
psaní mluvení zařazování

17. Podobně vypadající slova s odlišným významem (tzv. paronyma)	21. 7. 2017	ČJ počítač projektor slovní zásoba rodilý 8.-9. roč. SŠ porovnávání hledání	  
16. Typická substantiva v české poezii	21. 7. 2017	ČJ počítač projektor slovní zásoba stylistika rodilý 6.-7. roč. 8.-9. roč. SŠ hledání	  
15. Typická adjektiva v publicistice	21. 7. 2017	ČJ počítač projektor slovní zásoba stylistika rodilý 6.-7. roč. 8.-9. roč. SŠ hledání	  
14. Nejčastější slova	19. 6. 2017	ČJ počítač projektor slovní zásoba stylistika rodilý 6.-7. roč. 8.-9. roč. SŠ porovnávání hledání	  
13. Centrum a periferie slovní zásoby	5. 4. 2017	ČJ počítač projektor slovní zásoba rodilý 8.-9. roč. SŠ porovnávání hledání	  
12. Regionální varianty u tvarů slovesa <i>chtít</i>	5. 4. 2017	ČJ počítač projektor tvarosloví rodilý 6.-7. roč. 8.-9. roč. SŠ porovnávání hledání	  
11. Vykřičník v textech	5. 4. 2017	ČJ počítač projektor skladba stylistika rodilý 6.-7. roč. 8.-9. roč. SŠ hledání	  
10. Zdvojené souhlásky v českých slovech	2. 3. 2017	ČJ počítač projektor slovo tvorba rodilý 8.-9. roč. SŠ hledání	  
7. Dubletní psaní spřežek – dohromady, nebo zvlášť?	2. 3. 2017	ČJ počítač pravopis rodilý 8.-9. roč. SŠ hledání	  

 Pojdme se podívat na pár cvičení!



Máte jakékoli dotazy?

Sem s nimi!



v míru a prosperitě . Vážení přítomní ,	děkuji vám za pozornost	a přeji vám všechno dobré ! VOLBA PREZIDENTA
" S jeho slovy nelze než souhlasit .	Děkuji Vám za pozornost	. Co není rozbité , nemá se opravovat
prispívající k lepšímu životu na naší planetě .	Děkuji vám za pozornost	a přeji vašim jednáním úspěch . < P
spolupráci a navždy se s vámi rozloučil .	Děkuji vám za pozornost	. Pavel Zajíc ředitel pro koncepční a strategickou
nebude , pak je to rozhodně Senát .	Děkuji vám za pozornost	. Zápas o politický realismus Ať už si
abyste tuto svou vůli dali jasně najevo .	Děkuji vám za pozornost	i za důvěru , kterou mi stále projevujete
trvalou kuratelou OSN . Milí mladí přátelé !	Děkuji vám za pozornost	a nezapomeňte při odchodu vrátit papuče a zapsat

lucie.lukesova@korpus.cz



ČESKÝ NÁRODNÍ
KORPUS