

Úvod do korpusové lingvistiky 3



**DOSTUPNÉ KORPUSY A JEJICH STRUČNÁ
CHARAKTERISTIKA**

ÚČNK



- <http://ucnk.ff.cuni.cz/>
- Dostupné korpusy psaného jazyka
- Hledisko rozsahu a obsahu
- Hledisko anotací



Konference KL2014

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

Pracovní kolektiv

Pracovní kolektiv

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.

Korpusy psané češtiny



- Korpusy řady SYN
- Synchronní, psané, reprezentativní, obecné (žánrové zastoupení)
- Synchronní, psané, reprezentativní, specializované (PUB)

Korpusy psaného jazyka (synchronní)

korpus	velikost (počet slov)	lemmatizace	morfologické značky	rok zveřejnění	charakteristika korpusu
<u>SYN</u>	1 300 mil.	ANO	<u>ANO</u>	2010	<i>nereferenční</i>  spojení všech synchronních psaných korpusů řady SYN
<u>↑ SYN2010</u>	100 mil.	ANO	<u>ANO</u>	2010	žánrově vyvážený korpus, převažují texty z let 2005 - 2009
<u>↑ SYN2009PUB</u>	700 mil.	ANO	<u>ANO</u>	2010	korpus publicistických textů z let 1995 - 2007
<u>↑ SYN2006PUB</u>	300 mil.	ANO	<u>ANO</u>	2006	korpus publicistických textů z let 1989 - 2004
<u>↑ SYN2005</u>	100 mil.	ANO	<u>ANO</u>	2005	žánrově vyvážený korpus, převažují texty z let 2000 - 2004
<u>↑ SYN2000</u>	100 mil.	ANO	<u>ANO</u>	2000	žánrově vyvážený korpus, převažují texty z let 1990 - 1999
<u>FSC2000</u>	100 mil.	ANO	NE	2004	upravený SYN2000, referenční zdroj Frekvenčního slovníku češtiny
<u>CZESL-PLAIN</u>	2 mil.	NE	NE	2012	<i>nereferenční</i>  žákovský korpus češtiny nerodilých mluvčích
<u>KSK-DOPISY</u>	800 000	NE	NE	2006	přepisy ručně psané korespondence z let 1990 - 2004
<u>LINK</u>	1,8 mil.	ANO	<u>ANO</u>	2010	<i>nereferenční</i>  korpus sestavený z odborných lingvistických textů
<u>ORWELL</u>	80 000	ANO	<u>ANO</u>	2003	ručně označovaný korpus Orwellova románu "1984"
<u>SKRIPT2012</u>	590 000	ANO	<u>ANO</u>	2013	korpus školních písemných prací

Referenční – nereferenční korpus



- **Nereferenční korpus:** Většina korpusů ČNK jsou referenční entity, které zůstávají po celou dobu od svého zveřejnění neměnné, takže všechny dotazy, statistiky apod. jsou opakovatelné a dávají stále stejné výsledky. Některé korpusy však mají naopak **nereferenční** povahu, což znamená, že jsou průběžně vylepšovány a rozšiřovány. Všechny tyto změny jsou vždy po nějaké době promítnuty do již zveřejněného korpusu. K aktualizaci nereferenčního korpusu dochází nepravidelně, přibližně jednou ročně, většinou bez předchozího upozornění.

Synchronnost



- Hledisko produkce
- Hledisko recepce

Reprezentativnost



- Žánrové zastoupení v obecných korpusech řady SYN
- Zastoupení dle periodik v korpusech SYN_PUB

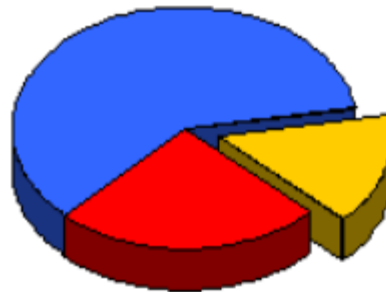
SYN2000

(100 milionů textových slov (tokens))



- Zastoupení žánrů

Složení korpusu SYN2000



60 % publicistika

25 % odborná literatura

15 % beletrie

SYN2005

(100 milionů textových slov (tokens))



- Zastoupení žánrů

Složení korpusu SYN2005:



40 % **beletrie**

27 % **odborná literatura**

33 % **publicistika**

SYN2010

(100 milionů textových slov (tokens))



- Zastoupení žánrů

Složení korpusu SYN2010:



40 % **beletrie**

27 % **odborná literatura**

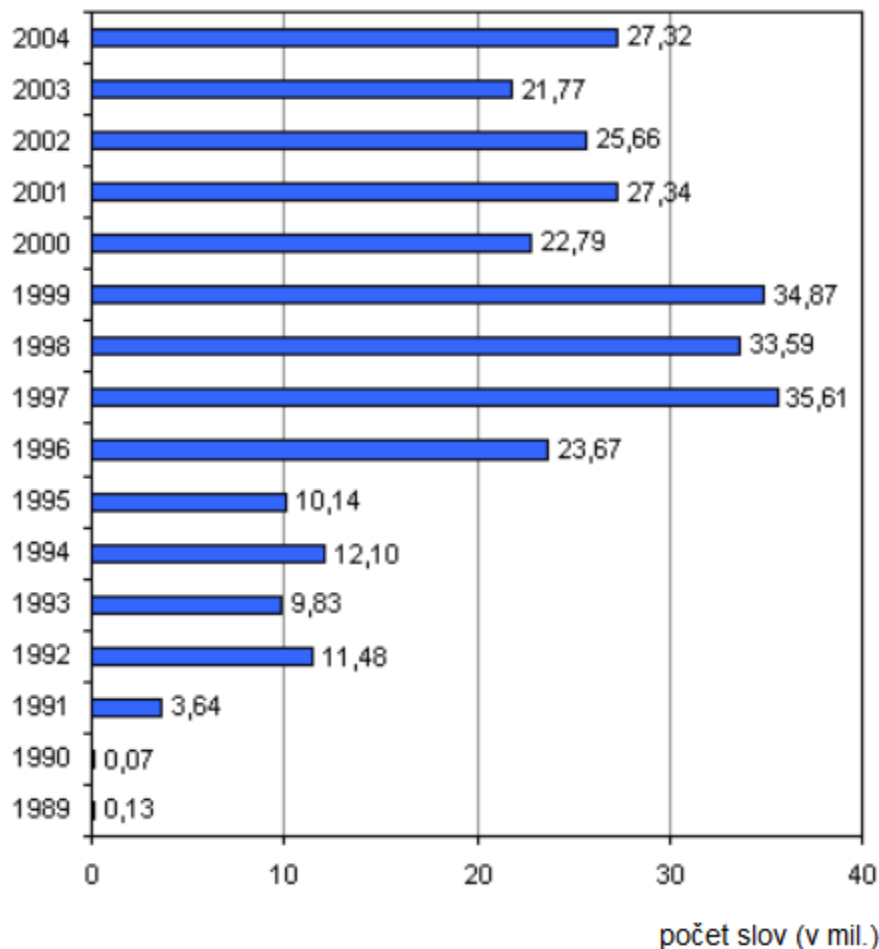
33 % **publicistika**

SYN2006PUB

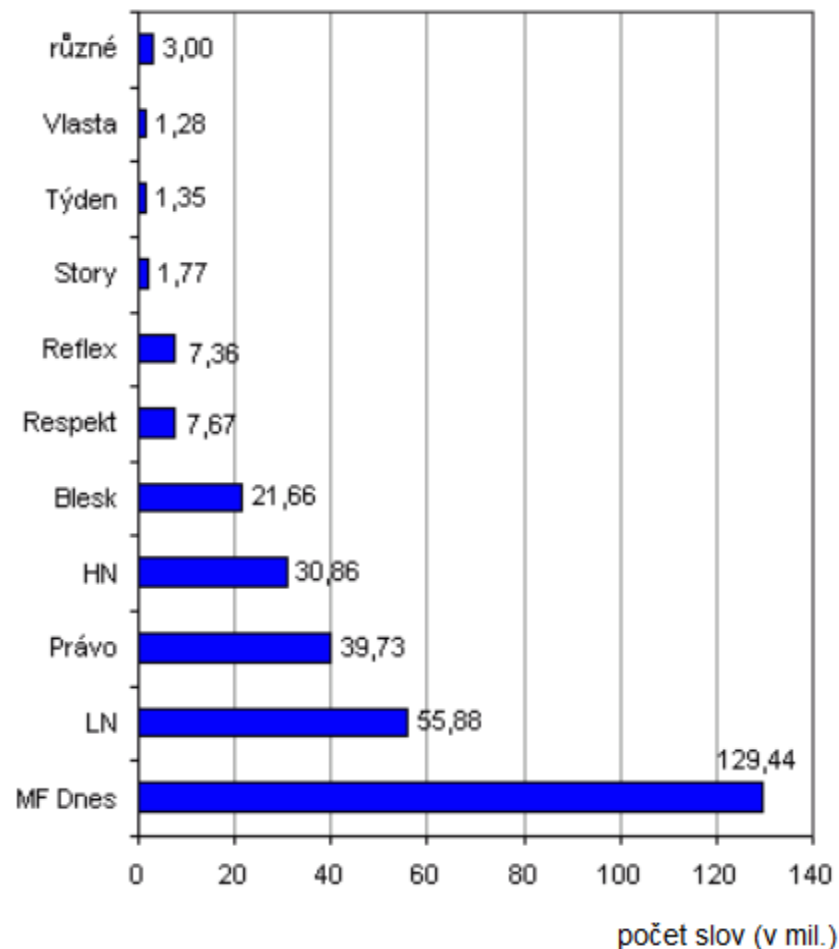
(300 milionů textových slov (tokens))



Složení korpusu podle roků



Složení korpusu podle titulů

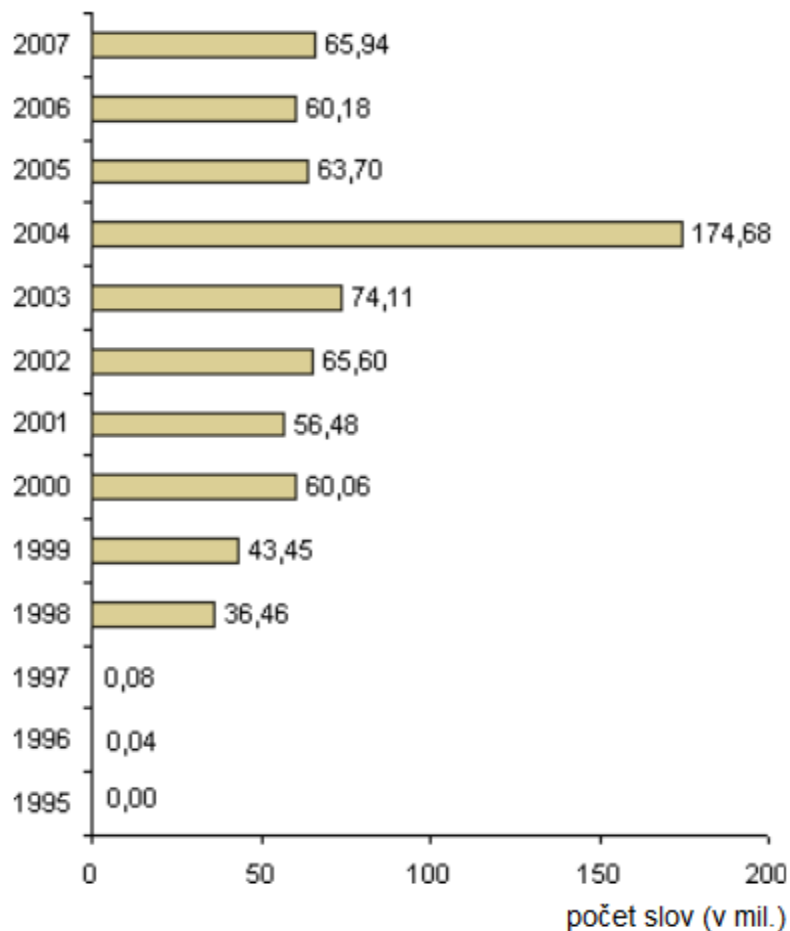


SYN2009PUB

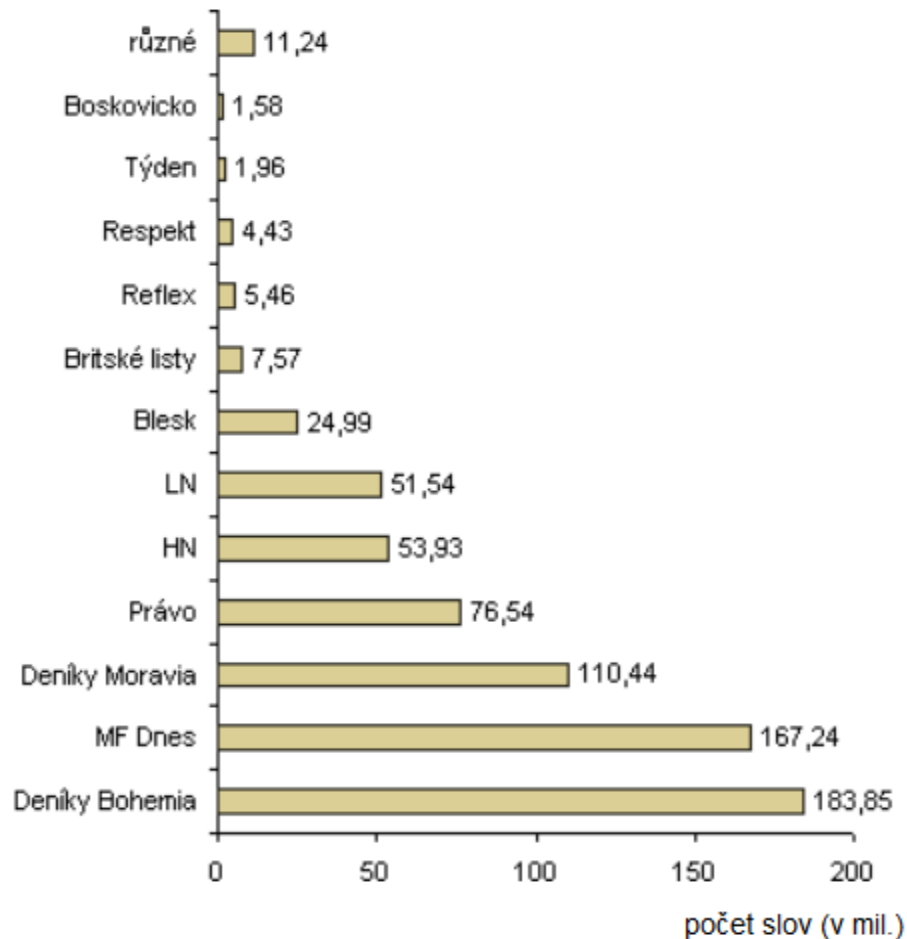
(700 milionů textových slov (tokens))



Složení korpusu podle roků



Složení korpusu podle titulů

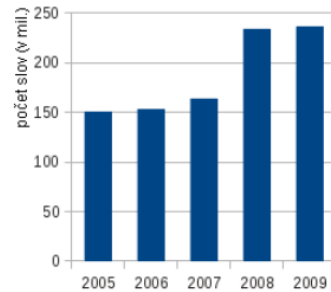


SYN2013PUB

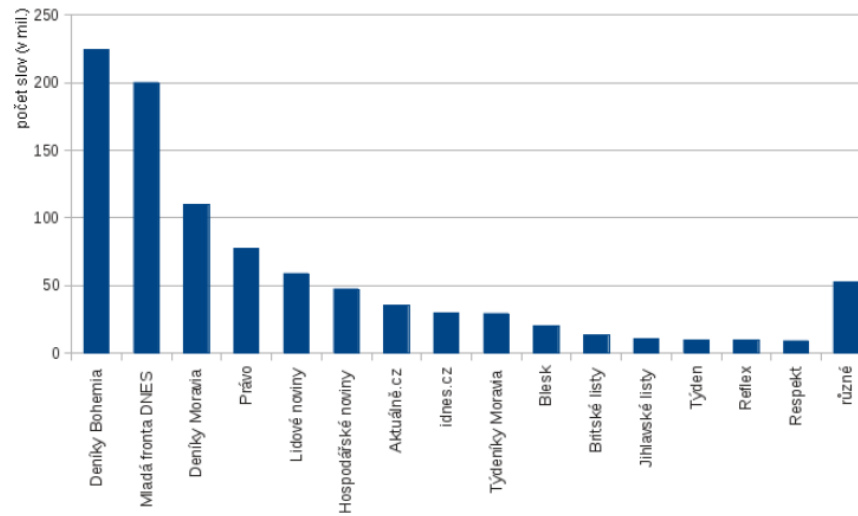
(935 milionů textových slov (tokens))



Složení korpusu podle roků



Složení korpusu podle titulů



Standardní anotace



- Vnětextová – kódy, možnost úplného zobrazení anotací
- Tokenizace
- Značkování odstavců
- Značkování vět
- Lemmatizace a morfologické značkování

Informace o rozsahu a obsahu korpusu při práci s korpusovým manažerem



← https://korpus.cz/corpora/run.cgi/first_form

Nejnavštěvovanější Jak začít

NoSketch Engine

Uživatel: osolsobe Korpus: [syn2010](#) Popis: [Synchronní reprezentativní korpus](#) Velikost: 121 667 413 pozic [?]

[Nový dotaz](#)
[Seznam slov](#) [?]

Pokročilá nastavení:
[Kontext](#) [?]
[Subkorpus](#) [?]

Uživatel:
[Změna hesla](#)

Korpus:

Typ dotazu:

Dotaz:

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

SYN2000



← <https://korpus.cz/corpora/run.cgi/first?reload=1&corpname=omezeni%2Fsyn2000&queryselector=iqueryrow&iqueryrow=1>

Nejnavštěvovanější Jak začít

NoSketch **Engin**

Uživatel: osolsobe Korpus:

[Nový dotaz](#)
[Seznam slov](#)

Pokročilá nastavení:
[Kontext](#)
[Subkorpus](#)

Uživatel:
[Změna hesla](#)

Korpus: syn2000

Atributy		Struktury	
word	1 763 813	doc	233 797
lemma	891 713	s	7 639 321
tag	2 019		
lc	1 462 975		
pos	12		
k	12		
g	11		
c	9		

Velikost: 120 908 724 pozic

KonText



Hledat v korpusu

Korpus:

syn2010

Typ dotazu:

Lemma:

Slovní druh: nespecifikováno

Specifikovat kontext

Specifikovat dotaz p

Hledat

Vymazat

▼ Synchronní psané korpusy

▼ řada SYN

[syn](#)

[syn2013pub](#)

[syn2010](#)

[syn2009pub](#)

[syn2006pub](#)

[syn2005](#)

[syn2000](#)

▶ specializované

▶ ke slovníkům

▶ Synchronní mluvené korpusy

▶ Diachronní korpusy

▶ Cizojazyčné korpusy

▶ Cizojazyčné korpusy webové

▶ Paralelní korpus InterCorp

korpus [syn2010](#)

121 667 413 pozic

Jak čítovat korpus

SYN2005



← <https://korpus.cz/corpora/run.cgi/first?reload=1&corpname=omezeni%2Fsyn2005&queryselector=iqueryrow&ique>

Nejnavštěvovanější Jak začít

NoSketch **Engin**

Uživatel: osolsobe Korpus:

[Nový dotaz](#)
[Seznam slov](#)

Pokročilá nastavení:
[Kontext](#)
[Subkorpus](#)

Uživatel:
[Změna hesla](#)

Korpus: syn2005

Atributy		Struktury	
word	1 778 142	opus	2 382
lemma	825 142	doc	132 353
tag	3 039	s	7 945 998
lc	1 475 902		
pos	12		
k	12		
g	11		
c	9		

Velikost: 122 419 382 pozic

SYN2010



← <https://korpus.cz/corpora/run.cgi/first?reload=1&corpname=omezeni%2Fsyn2010&queryselector=iqueryrow&iq>

Nejnavštěvovanější Jak začít

NoSketch **Engin**

Uživatel: osolsobe Korpus:

[Nový dotaz](#)
[Seznam slov](#)

Pokročilá nastavení:
[Kontext](#)
[Subkorpus](#)

Uživatel:
[Změna hesla](#)

Korpus: syn2010

Atributy		Struktury	
word	1 706 345	opus	2 649
lemma	785 580	doc	152 634
tag	4 289	s	8 172 649
lc	1 406 404		
pos	12		
k	12		
g	6		
c	8		

Velikost: 121 667 413 pozic

KonText



syn2010

velikost: 121 667 413 pozice

webová stránka: <http://www.korpus.cz/syn2010.php>

Atributy		Struktury	
word	1 706 345	<opus>	2 649
lemma	785 580	<doc>	152 634
tag	4 289	<s>	8 172 649
lc	1 406 404		
pos	12		
k	12		
g	6		
c	8		

Poznámka: čísla v tomto přehledu udávají počty různých atributů / struktur (tj. typů) v korpusu.

Synchronní reprezentativní korpus

syn2013pub

velikost: 1 120 014 835 pozice

webová stránka: <http://www.korpus.cz/syn2013pub.php>

Atributy		Struktury	
word	4 200 464	<opus>	21 469
lemma	2 549 185	<doc>	4 172 882
tag	4 392	<s>	76 681 361
lc	3 499 046		
pos	12		
k	12		
g	5		
c	8		

Poznámka: čísla v tomto přehledu udávají počty různých atributů / struktur (tj. typů) v korpusu.

Synchronní publicistický korpus

Definice word (tokenizace <http://wiki.korpus.cz/doku.php/pojmy:token>)



- Řetězec znaků mezi oddělovači
- Problémy tokenizace
- Když jedné jednotce na úrovni systému odpovídá více jednotek na úrovni textu a naopak

Lemma (<http://wiki.korpus.cz/doku.php/pojmy:lemma>)



- Textové slovo – systémové slovo
- Reprezentativní tvar
- Lemmatizace prováděná pomocí automatických nástrojů
- Lemma = tvar sám

Tag



- Tagset
- Poziční systém
- Atribut/hodnota
- Klasické gramatické kategorie a morfologické tagy

<http://wiki.korpus.cz/doku.php/seznamy:tagy>



Příručka uživatele

Internetová příručka uživatele korpusů ČNK obsahuje ve formě wiki přehledný návod pro práci s korpusem doplněný o slovníček lingvistických termínů, seznam korpusů ČNK spolu s jejich specifikací a další užitečné informace.

Obsah

- Morfologické značky (tagy)
 - Struktura značky
 - Změny v morfologickém značkování
 - Popis jednotlivých pozic značky
 - Pozice 1 - Slovní druh
 - Pozice 2 - Detailní určení slovního druhu
 - Pozice 3 - Jmenný rod
 - Pozice 4 - Číslo
 - Pozice 5 - Pád
 - Pozice 6 - Přivlastňovací rod
 - Pozice 7 - Přivlastňovací číslo
 - Pozice 8 - Osoba
 - Pozice 9 - Čas
 - Pozice 10 - Stupeň
 - Pozice 11 - Negace
 - Pozice 12 - Aktivum/pasivum
 - Pozice 13 - Nepoužito
 - Pozice 14 - Nepoužito
 - Pozice 15 - Varianta (stylový příznak)
 - Pozice 16 - Vid

Morfologické značky (tagy)



- Morfologické značky (tagy) jsou součástí výsledku (výstupem) morfologické analýzy, která pracuje s izolovanými slovními tvary, tedy bez ohledu na jejich kontext. Druhou částí výsledku je tzv. lemma, které identifikuje příslušnou lexikální jednotku ve smyslu slovníkového hesla. Morfologická analýza je obecně nejednoznačná; slovní tvary, brány izolovaně a bez ohledu na kontext, pochopitelně nemohou být v mnoha případech jednoznačně určeny, a to jak z hlediska lemmatu, tak z hlediska morfologické značky. V druhé fázi dochází k desambiguaci (zjednoznačnění), která z plejády možných interpretací vybírá v ideálním případě tu nejvhodnější.



- Morfologické značky slouží k snadnějšímu hledání v korpusech (povětšinou pouze [psané češtiny](#)), jejich účelem tedy není být základem pro analýzu konkrétních výskytů. Automatická analýza není přirozeně bezchybná, podíl špatně určených značek se odhaduje na 4 % (úroveň kolísá v závislosti na typu morfologické kategorie).

Large web corpora



Corpora

Language	Corpus name	Tokens	Words	
Czech	czes2	465,102,710	350,990,047	
Czech	czTenTen12 [cleaned]	5,214,920,358	4,291,223,836	
Czech	desam	1,042,973	874,354	
English	British Academic Spoken English Corpus (BASE)	1,252,256	1,186,290	
English	British Academic Written English Corpus (BAWE)	8,336,262	6,964,411	
English	British National Corpus	112,181,015	96,048,950	
English	Susanne	150,426	128,998	
English	ukWaC	1,565,274,190	1,318,047,961	

[Show all corpora](#) | [Parallel corpora](#)

My corpora

Language	Corpus name	Configuration template	Tokens	
Czech	sheltie	Desamb for Czech	31,141	
English	Apple	TreeTagger for English	52,253	

[Create corpus](#) | [WebBootCaT](#)

czTenTen12 (5,5 miliard tokenů)



czTenTen12

Czech [TenTen family](#) web corpus crawled by [SpiderLing?](#) in 2011 and Heritrix in 2010. Encoded in UTF-8, cleaned, deduplicated. Tagged by Desamb in 2012.

Changelog

v1 (April 2012)

- initial version -- 4.8 G words

v2 (September 2012)

- tagged by Ajka + Desamb

v3 (December 2012)

- retagged, corrected
- word sketches
- still not a final version