CZECH NATIONAL CORPUS

Recent Developments in the Czech National Corpus

Michal Křen
Charles University in Prague

3rd Workshop on the Challenges in the Management of Large Corpora
Lancaster
20 July 2015

Introduction of the project

Corpus compilation
- Written corpora
- Spoken corpora
- Parallel corpus
- Specialized corpora

Data processing and annotation
- Project management tools
- Tools for linguistic annotation

User application development
- KonText
- SyD, Morfio & KWords

Services
- Wiki, Support & Biblio
- Corpus hosting
- Data packages

Future plans

Introduction of the project

- Corpus compilation
  - Written corpora
  - Spoken corpora
  - Parallel corpus
  - Specialized corpora
- Data processing and annotation
  - Project management tools
  - Tools for linguistic annotation
- User application development
  - KonText
  - SyD, Morfio & KWords
- Services
  - Wiki, Support & Biblio
  - Corpus hosting
  - Data packages
- Future plans

Czech National Corpus

- long-term project (since 1994)
- continuous mapping of Czech language
- compilation, maintenance and providing public access to various language corpora
- research infrastructure (since 2012) $\Rightarrow$ service-oriented operation
- more than 4,500 registered active users
- almost 1,900 queries a day
- `http://www.korpus.cz`

corpus	*size*	*contents*	*time span*
SYN2000	100 mil.	representative	most of the texts from 1900–1999
SYN2005	100 mil.	representative	most of the texts from 2000–2004
SYN2010	100 mil.	representative	most of the texts from 2005–2009
SYN2006PUB	300 mil.	newspaper	1989–2004
SYN2009PUB	700 mil.	newspaper	1995–2007
SYN2013PUB	935 mil.	newspaper	2005–2009
SYN (version 3)	2 232 mil.	union	

Currently available SYN-series corpora.

- traditional corpora with detailed bibliographical information
- lemmatized & morphologically tagged

corpus	*size*	*contents*	*time span*
SYN2000	100 mil.	representative	most of the texts from 1900–1999
SYN2005	100 mil.	representative	most of the texts from 2000–2004
SYN2010	100 mil.	representative	most of the texts from 2005–2009
SYN2006PUB	300 mil.	newspaper	1989–2004
SYN2009PUB	700 mil.	newspaper	1995–2007
SYN2013PUB	935 mil.	newspaper	2005–2009
SYN (version 3)	2 232 mil.	union	

Currently available SYN-series corpora.

- traditional corpora with detailed bibliographical information
- lemmatized & morphologically tagged
- outlook:
  - new representative corpus SYN2015
  - fresh data in SYN (2010–2014 added)

corpus	*size*	*coverage*	*time span*
ORAL2006	1 mil.	Bohemia	recordings from 2002–2006
ORAL2008	1 mil.	Bohemia	recordings from 2002–2007
ORAL2013	2.78 mil.	Czech Republic	recordings from 2008–2011

Currently available ORAL-series corpora.

- only unscripted, informal dialogical speech
- ORAL2013 designed as a representation of contemporary spontaneous spoken Czech
- manual one-layer transcription

corpus	*size*	*coverage*	*time span*
ORAL2006	1 mil.	Bohemia	recordings from 2002–2006
ORAL2008	1 mil.	Bohemia	recordings from 2002–2007
ORAL2013	2.78 mil.	Czech Republic	recordings from 2008–2011

Currently available ORAL-series corpora.

- only unscripted, informal dialogical speech
- ORAL2013 designed as a representation of contemporary spontaneous spoken Czech
- manual one-layer transcription
- outlook:
  - lemmatization & tagging
  - two-layer ORTOFON series

InterCorp

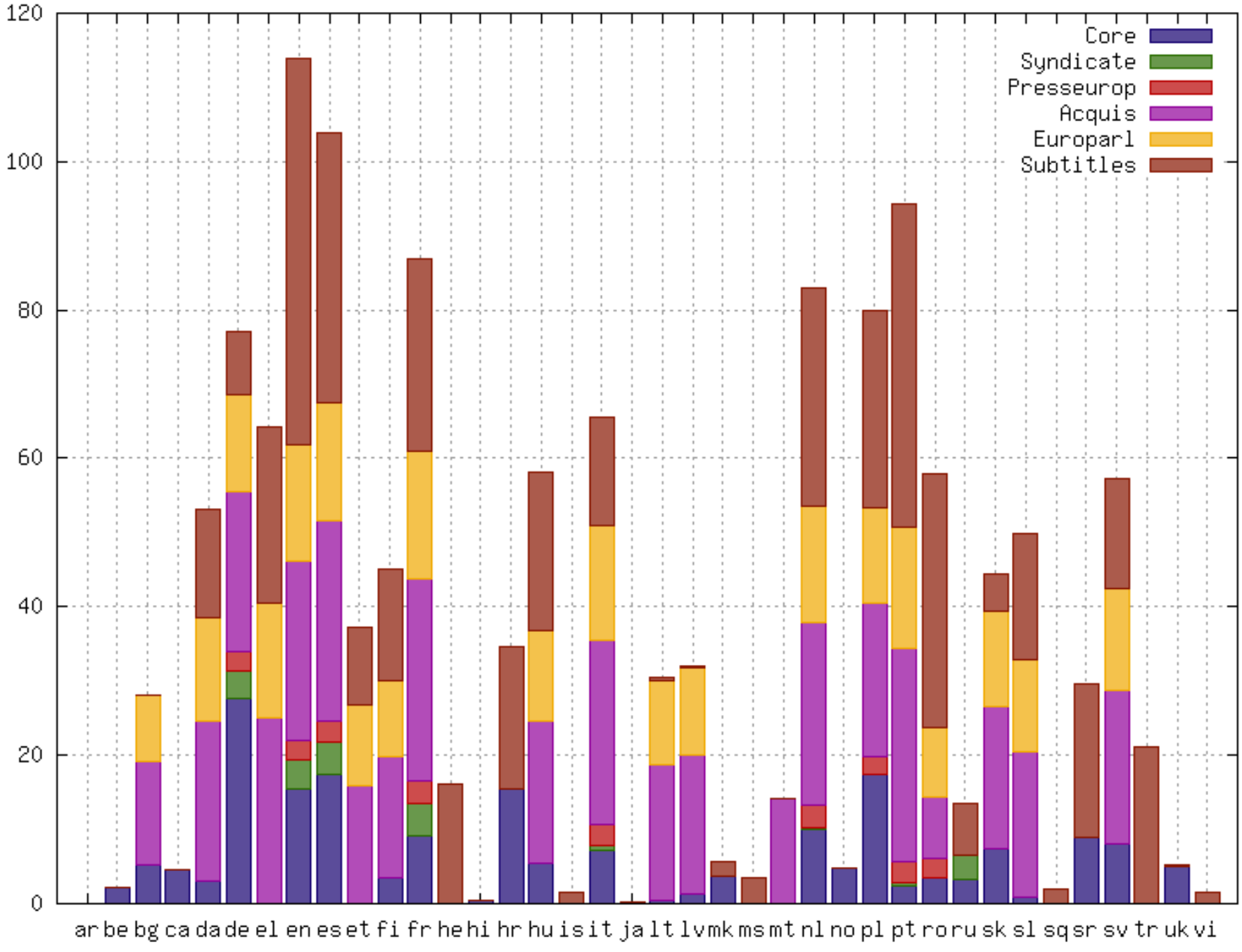
- large parallel corpus
- texts aligned on sentence level with their translations between Czech and a number of other languages
- consists of two major parts:
  - ***core***: manually revised alignment, mostly fiction
  - ***collections***: automatic alignment, various domains

InterCorp

- large parallel corpus
- texts aligned on sentence level with their translations between Czech and a number of other languages
- consists of two major parts:
  - ***core***: manually revised alignment, mostly fiction
  - ***collections***: automatic alignment, various domains

Version 8 (June 2015)

- 38 foreign languages, out of which 20 lemmatized and/or tagged
- foreign-language texts:
  size of the core: 194 mil., total size of the InterCorp: 1,423 mil. words
- collections included:
  - journalistic texts: Project Syndicate, Presseurop
  - Acquis Communautaire, EuroParl, Open Subtitles


120
100
80
60
40
20
0
Core
Syndicate
Presseurop
Acquis
Europarl
Subtitles
ar be bg ca da de el en es et fi fr he hi hr hu is it ja lt lv mk ms mt nl no pl pt ro ru sk sl sq sr sv tr uk vi

DIAKORP

- diachronic corpus of historical Czech (from 14^{th} century onwards, with current focus on the 19^{th} century)
- current size 2 mil. words, major update soon

DIAKORP

- diachronic corpus of historical Czech (from 14^{th} century onwards, with current focus on the 19^{th} century)
- current size 2 mil. words, major update soon

DIALEKT

- dialectal corpus
- target size 200,000 words (end of 2016)

DIAKORP

- diachronic corpus of historical Czech (from 14^{th} century onwards, with current focus on the 19^{th} century)
- current size 2 mil. words, major update soon

DIALEKT

- dialectal corpus
- target size 200,000 words (end of 2016)

DEAF

- corpus of Czech texts written by the deaf
- target size 200,000 words (end of 2016)

Project management tools

- software environments for internal work flow management
- web-based “wrappers” that combine both CNC and third-party tools

SynKorp

- database interconnected with data processing toolchain
- data collection and processing of the written language corpora
  - customizable text conversion and clean-up
  - bibliographic annotation and text classification

Mluvka

- database and integrated project management system
- coordination of spoken and dialectal data collection
- large networks of external collaborators
  $\Rightarrow$ three-level project coordination hierarchy
  - manual two-layer annotation (orthographic and phonetic)
  - formal compliance checks and expert revisions
  - balancing of the collected material
  - payment calculation

InterCorp database

- database and integrated project management system
- coordination of data collection for InterCorp
- large networks of external collaborators
  $\Rightarrow$ three-level project coordination hierarchy
  - work flow management of the individual texts
  - manual verification and revision of the alignment (using InterText, a project-independent editor of aligned parallel texts)
  - payment calculation

Tools for linguistic annotation

- morphological and syntactic level
- typical model: Czech-specific tools built upon language-independent third-party ones

Morphological level

- morphological analysis
  - continuous CNC feedback to the dictionary provided by LINDAT/CLARIN
- disambiguation
  - third-party stochastic tagger
  - rule-based components developed by the CNC

Morphological level

- morphological analysis
  - continuous CNC feedback to the dictionary provided by LINDAT/CLARIN
- disambiguation
  - third-party stochastic tagger
  - rule-based components developed by the CNC

Syntactic level

- dependency parsing
  - third-party stochastic parser
  - rule-based corrections and other enhancement methods

CZECH NATIONAL CORPUS

Type a word or phrase

Search the corpus

ignore case search word forms of the given lemma

WEB CORPORA ARANEA

JUNE 18, 2015

A family of non-reference comparable web corpora called Aranea and compiled by Vladimír Benko has been updated in June 2015. It currently covers 14 languages (cs, de, en, es, fi, fr, hu, it, nl, pl, pt, ru, sk, zh).

INTERCORP RELEASE 8

JUNE 4, 2015

A new release of the InterCorp parallel corpus has been published in June 2015. The total size of foreign language texts has reached 194 million tokens in the core and 1.2 billion tokens in collections.

TERMINATION OF OLDER INTERFACES

APRIL 8, 2015

As of the last of March, support for the Bonito, Park and NoSkE applications was officially ended. Users of these applications can ease their transition to the new KonText interface with a set of useful tips prepared specially for the occasion.

For more detailed information continue here.

What is a corpus?

A language corpus is an electronic collection of authentic texts (written or spoken) easily searchable for various language phenomena (esp. words and collocations) and to display them in their natural context.

more...

Who are we?

CZECH NATIONAL CORPUS

The Czech National Corpus is an **academic project** founded in 1994 at the CU FA and administered by the Institute of the Czech National Corpus. The aim of the project is systematic mapping of Czech and other languages in comparison with Czech. CNC corpora are accessible to everybody interested in studying the language after free registration.

more...

Applications

kontext

The KonText application is a **basic query interface for**

Support and information resources

KonText

- `http://kontext.korpus.cz`
- web-based general-purpose corpus concordancer
- CNC fork of the NoSketch Engine
- single interface to all corpus types (including parallel and spoken corpora)
- built-in basic statistical functions, subcorpus manager, filtering etc.
- requires user registration to switch from restricted functionality

Hits: **154** | i.p.m.: 1.10 (related to the whole InterCorp v8 - English) | ARF: **19.88** | Result is shuffled 1 / 4

The Wonderful Adventures of Nils	thought no more of housemaids and men , who were	hounding	him , but climbed up the steps - and into
Between the Acts	as well as Buster ? " </s><s> Colin was his famous	hound	. </s><s> But there was only room for Buster . </s><s> It
The fellowship of the Ring	we are coming with you ; </s><s> or following you like	hounds	. ' </s> <s> 'And after all , sir , ' added
The Master and Margarita	, remarkable even for a theatre manager . </s><s> The free-ticket	hounds	, for instance , regarded him as their patron saint
The Testament	should be my choice too , but I 'm being	hounded	. </s> <s> Why should I care who gets the money ? </s>
Lovers and Murderers	and fewer honors . </s><s> All the same ... a hundred	hounds	for every stag . </s><s> When Neustupa takes my place in
Hope Abandoned	groups as well - before he was himself dismissed while	hounding	Professor Lubishchev , a biologist who had once spoken up
English Fairy Tales	let the young prince in with his horse and his	hound	" , and Kate added , " and his lady
The Street Lawyer	<s> He had a lot of bills . </s><s> Credit agencies were	hounding	him . </s><s> For the moment , he was hiding at
Hope Abandoned	thJ spirit in which Akhmatova , Mandelstam and Pasternak were	hounded	- and Mayakovski , too , until he was made
A Song of Stone	picking up the dead birds drops , obedient as any	hound	. </s><s> Another flock of birds is circling , curving down-slope
The Encyclopedia of the Dead	Mariette 's funeral ; dogs barked all night ; the	hounds	were called out , and Alsatians straining at the collar
Among the Bears	could have been run out of the area by the	hounds	or just been behind a ridge from me . </s> <s> On
The Client	. </s><s> Because if we do , a million cops 'll	hound	us to our graves . </s><s> It wo n't work .
The Master and Margarita	refuge in the treasurer 's office from the complimentary ticket	hounds	who made his life a misery , especially on the
Harry Potter and the Order of the Phoenix	eyes that gave him the doleful look of a basset	hound	. </s><s> He was also clutching a silvery bundle that Harry
Hope Abandoned	would have been far worse if , instead of being	hounded	, they had flourished and Gorodetski had made himself their
The Silence of the Lambs	<s> But not without casualties . </s><s> Will Graham , the keenest	hound	ever to run in Crawford 's pack , was a
The Virgin and the Gypsy	knees . </s><s> She pretended to be interested in the white-and-liver-coloured	hound	. </s> <s> ' How much do you want , if we
Lolita	paradise loose . </s><s> I had ceased to be Humbert the	Hound	, the sad-eyed degenerate cur clasping the boot that would
Between the Acts	brute like that obey him . </s><s> Back came the Afghan	hound	, sidling , apologetic . </s><s> And as he cringed at
Single & Single	know , he 's cringing in a ditch with the	hounds	after him and all I want to do is reach
Dictionary of the Khazars	itself was not a good omen . </s><s> They watched their	hounds	race through the scents of the Bosnian woods as through
The two towers	! ' </s> <s> They went in single file , running like	hounds	on a strong scent , and an eager light was
Good Omens	re you goin ' to call it ? ' </s> <s> The	hound	waited . </s><s> This was the moment . </s><s> The Naming . </s>
The open society and its enemies	to taste blood ; </s><s> just as one does with young	hounds	' . </s><s> The description of a modern writer , who

SyD

- `http://syd.korpus.cz`
- corpus-based analysis of language variants
- synchronic and diachronic component
- synchronic comparison of frequency distribution of variants across different domains of contemporary written and spoken texts
- diachronic development over time
- available without registration

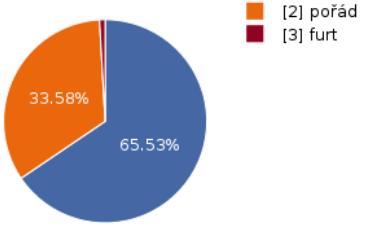
SyD

Diachronní Synchronní Nový dotaz

[1] stále [2] pořád [3] furt

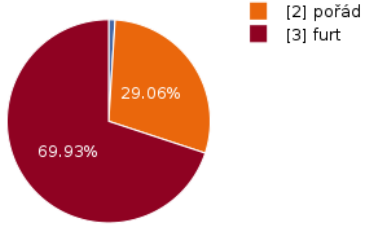
Porovnat varianty a=A lemma Zobrazit adresu dotazu Sdílet

Psaný jazyk



Uložit jako

Mluvený jazyk



Uložit jako

Celkové údaje pro psaný a mluvený jazyk (které můžou být vzhledem k velké obecnosti zkreslující, viz **Rozložení**). Přesné údaje o poměrech se zobrazí po kliknutí na libovolnou oblast grafu.

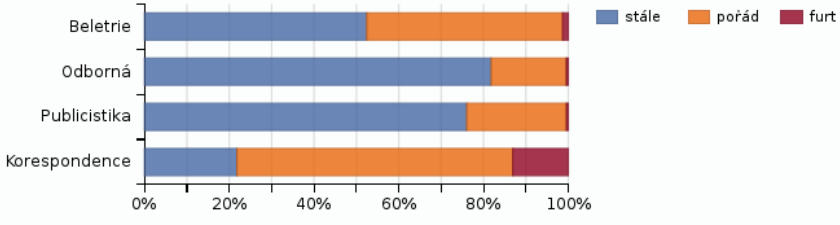
Údaj *Nedostatečná data* značí, že součet frekvencí všech variant v daném (sub)korpusu je menší než 5 výskytů.

Zadání dotazu a jeho výsledek je uložen na serveru a může být znovu vyvolán pomocí odkazu uvedeného v záhlaví stránky. Odkaz je možné použít i pro citační účely.

Souhrn Rozložení Psaný jazyk Mluvený jazyk Kolokace

Typy textů (základní dělení)

Typy textů: souhrn



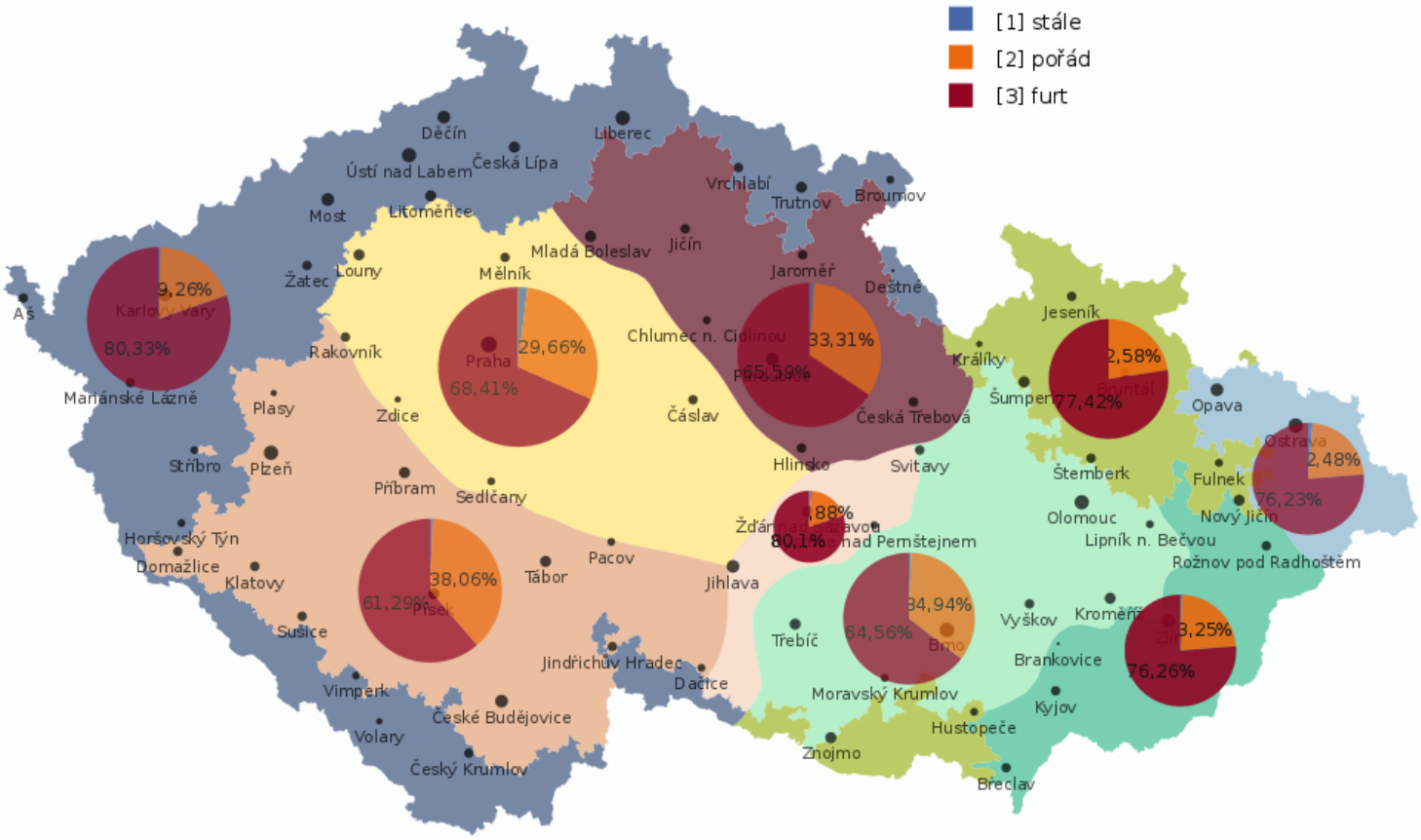
Uložit jako

Graf ukazuje poměr distribuce variant v základních typech textů (v odborné terminologii *registr*, v datové struktuře atribut *txtypegroup*): beletrii, publicistice a odborné literatuře (údaje pocházejí z korpusu SYN2010). Jako další makroskupina textů je zde přidána korespondence, která představuje psaný, ovšem neveřejný a neoficiální typ textu, a kterou odráží korpus KSK-Dopisy.

Po kliknutí na libovolnou oblast grafu se zobrazí tabulka s číselnými údaji pro každou z variant: absolutní frekvence (tj. počet výskytů dané varianty v konkrétním subkorpusu), relativní frekvence v *ppm* (počet výskytů na milion slov), která umožňuje srovnání napříč jednotlivými kategoriemi textů navzdory jejich nestejné velikosti, a konečně relativní frekvence v procentech, vyjadřující srovnání s ostatními variantami (součet této hodnoty u všech variant se rovná 100%).

Nářeční oblasti

Oblast: českomoravská



SyDiK

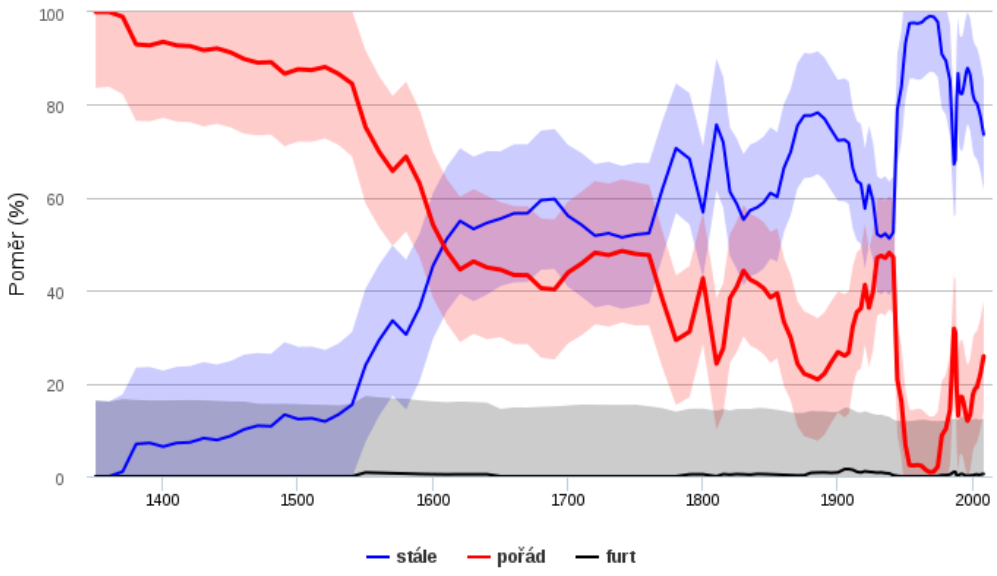
Diachronní | Synchronní

Nový dotaz

[1] stále [2] pořád [3] furt

Okno: proměnlivé Krok: proměnlivý

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. Uživatele 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.

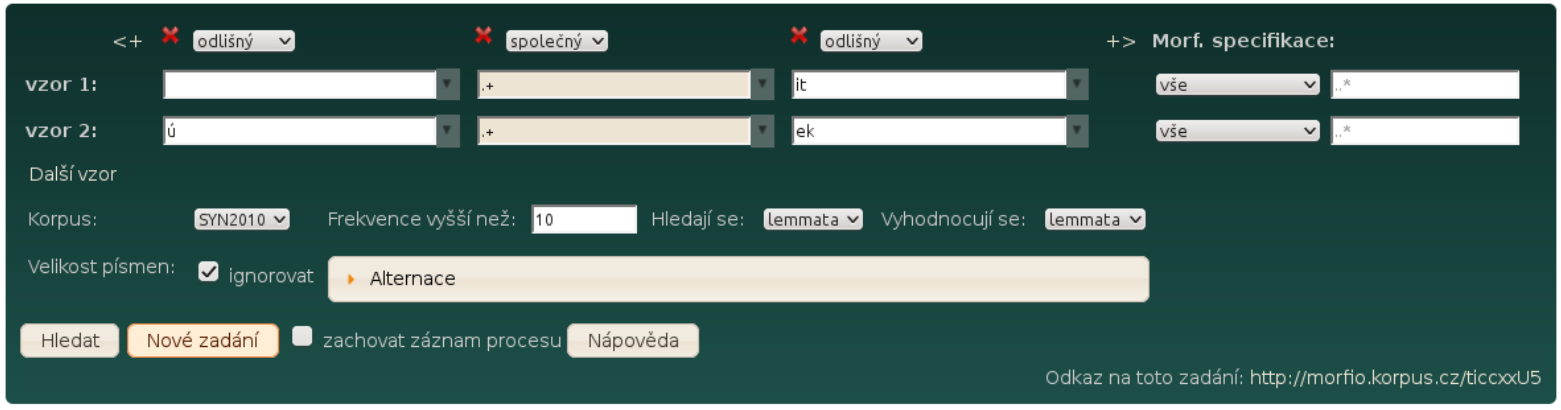
Úvodní graf prezentuje tendence vývoje poměru variant v průběhu času. Každá linie reprezentující jednu variantu je obklopena stejnobarevnou oblastí, jejíž šíře značí spolehlivost naměřených dat (čím širší oblast, tím větší možná **chybovost** výsledku).

více

Morfio

- `http://morfio.korpus.cz`
- word formation and derivational morphology
- identifies selected derivational patterns specified by affixes and word roots
- analysis of their morphological productivity
- available without registration

Morfio



Souhrn | Výpis | Produktivita | vzor 1 | vzor 2

	vzor 1 (fq)	vzor 2 (fq)
1	činit (10887)	účinek (7236)
2	lomit (202)	úlomek (790)
3	lovit (2070)	úlovek (682)
4	lit (17)	úlek (361)
5	platit (30275)	úplatek (1513)
6	radit (3999)	úradek (57)
7	sit (49)	úsek (7292)
8	šklebit (452)	úšklebek (561)
9	tržit (13)	útržek (781)
10	tit (46)	útek (27)
11	tulit (128)	útulek (1136)

V tabulce jsou uvedeny všechny doklady ze všech vzorů, které vstupují do zadaného modelu. Červená část slov označuje společnou bázi (ta se může lišit pouze v případě aplikace alternací). V závorkách uvedený údaj představuje celkovou frekvenci jednotky ve zvoleném korpusu. Tabulku je možné přetřídit podle libovolného sloupce a to jak abecedně, tak frekvenčně pomocí šipek v záhlaví tabulky. Každé slovo zároveň funguje jako odkaz směřující k ukázce konkordancí ve zvoleném korpusu.

Páry vytvořené až díky aplikaci alternačních pravidel jsou zvýrazněné barevným pozadím. Jejich báze se proto budou lišit. V případě, že použitím alternačních pravidel dojde k situaci, že jednomu vzoru v dané dvojci odpovídá více slov, jsou všechna tato slova uvedena hromadně v jednom řádku tabulky.

KWords

- `http://kwords.korpus.cz`
- corpus-based keyword and discourse analysis
- possibility to upload texts to be analyzed and/or the reference text
- visualization of distance-based keyword relations
- available without registration

IKI words

Jazyk vstupního textu: Angličtina

Nápověda

Vstupní text (který chcete analyzovat):

- Vložte text
- Nahrajte textové soubory / Multi-analýza

Referenční korpus (úzus, s nímž chcete text porovnávat):

- Vyberte referenční korpus
  - BNC
- Vložte referenční text
- Nahrajte referenční textový soubor

Stop-list (slova, která mají být vyloučena z analýzy):

- [x] zájmena
- [x] předložky
- [x] spojky
- [x] čísla

Nastavení

Velikost písmen:

- [x] ignorovat

Analyzovat

Text | Klíčová slova | Distribuce | Keyword links | Konkordance

14

THE HOUND OF THE BASKERVILLES

One of Sherlock Holmes's defects - if, indeed, one may call it a defect - was that he was exceedingly loth to communicate his full plans to any other person until the instant of their fulfilment. Partly it came no doubt from his own masterful nature, which loved to dominate and surprise those who were around him. Partly also from his professional caution, which urged him never to take any chances. The result, however, was very trying for those who were acting as his agents and assistants. I had often suffered under it, but never more so than during that long drive in the darkness. The great ordeal was in front of us; at last we were about to make our final effort, and yet Holmes had said nothing, and I could only surmise what his course of action would be. My nerves thrilled with anticipation when at last the cold wind upon our faces and the dark, void spaces on either side of the narrow road told me that we were back upon the moor once again. Every stride of the horses and every turn of the wheels was taking us nearer to our supreme adventure.

Our conversation was hampered by the presence of the driver of the hired wagonette, so that we were forced to talk of trivial matters when our nerves were tense with emotion and anticipation. It was a relief to me, after that unnatural restraint, when we at last passed Frankland's house and knew that we were drawing near to the Hall and to the scene of action. We did not drive up to the door, but got down near the gate of the avenue. The wagonette was paid off and ordered to return to Coombe Tracey forthwith, while we started to walk to Merripit House.

"Are you armed, Lestrade?"

The little detective smiled. "As long as I have my trousers, I have a hip-pocket, and as long as I have my hip-pocket I have something in it."

"Good! My friend and I are also ready for emergencies."

"You're mighty close about this affair, Mr Holmes. What's the game now?"

"A waiting game."

CNC Wiki

- `http://wiki.korpus.cz`
- description of corpora available
- reference manual of KonText including a tutorial in 7 lessons
- introduction to corpus linguistics
- available without registration

User Forum

- `http://podpora.korpus.cz`
- advisory centre with Q&A
- bug reporting
- requests for new features
- available only to registered users

Biblio

- `http://biblio.korpus.cz`
- repository of CNC-based research outputs
- references and/or uploaded full papers
- available without registration
- motivation:
  - bibliography of Czech corpus linguistics
  - promoting Open Access
  - promoting individual papers
  - helping the CNC project

Corpus hosting

- service offered to other research groups consisting in:
  - final technical processing of their corpus data
  - (possibly extensive) consistency checks
  - publication, maintenance, public access and related services
- mutual advantage, credit always given
- examples:
  - CzeSL series of Czech learner corpora
    (by Karel Šebesta et al.)
  - DOTKO and HOTKO (Lower and Upper Sorbian, respectively;
    by Sorbian Institute, Bautzen, Germany)
  - Aranea series of large comparable web corpora
    (currently 14 languages; by Vladimír Benko)

Data packages

- providing corpus-derived data with less restrictive licensing
- offered to users who need direct access to the corpus data (NLP)
- availability:
  - LINDAT/CLARIN repository ("standard packages")
  - on demand, in accordance with individual requirements
- licensing depends on the nature of the data and it ranges between:
  - CC-BY (e.g. word lists or n-grams for small n)
  - restrictive proprietary license that permits neither commercial use nor redistribution (e.g. full texts shuffled at sentence level)

User applications

- continuous maintenance, adding new functionality
- KonText enhancements:
  - module leading the users to appropriate statistical evaluation and interpretation of the results
  - multidimensional frequency distribution with attractive visualisation (e.g. diachronic development)
  - advanced collocational component
  - subcorpus blending module

User applications

- continuous maintenance, adding new functionality
- KonText enhancements:
  - module leading the users to appropriate statistical evaluation and interpretation of the results
  - multidimensional frequency distribution with attractive visualisation (e.g. diachronic development)
  - advanced collocational component
  - subcorpus blending module

Data collection

- semi-formal spoken Czech
- semi-official internet language (blogs, discussion forums etc.)
- monitor corpus of written Czech (1850–present)

Thank you for your attention!

Selected references

Čermák, F. – Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 13 (3), 411–427.

Hnátková, M. – Křen, M. – Procházka, P. – Skoumalová, H. (2014). The SYN-series corpora of written Czech. In: *Proceedings of LREC 2014*, 160–164. Reykjavík: ELRA.
http://www.lrec-conf.org/proceedings/lrec2014/pdf/294_Paper.pdf

Jelínek, T. (2014). Improvements to Dependency Parsing Using Automatic Simplification of Data. In: *Proceedings of LREC 2014*, 73–77. Reykjavík: ELRA.
http://www.lrec-conf.org/proceedings/lrec2014/pdf/228_Paper.pdf

Kopřivová, M. – Goláňová, H. – Klimešová, P. – Lukeš, D. (2014). Mapping Diatopic and Diachronic Variation in Spoken Czech: the Ortofon and Dialekt Corpora. In: *Proceedings of LREC 2014*, 376–382. Reykjavík: ELRA.
http://www.lrec-conf.org/proceedings/lrec2014/pdf/252_Paper.pdf

Kučera, K. – Stluka, M. (2014). Corpus of 19th-century Czech Texts: Problems and Solutions. In: *Proceedings of LREC 2014*, 165–168. Reykjavík: ELRA.
http://www.lrec-conf.org/proceedings/lrec2014/pdf/300_Paper.pdf

Petkevič, P. (2006). Reliable Morphological Disambiguation of Czech: Rule-Based Approach is Necessary. In: *Insight into Slovak and Czech Corpus Linguistics*, 26–44. Bratislava: Veda.

Rosen, A. – Vavřín, M. (2012). Building a multilingual parallel corpus for human users. In: *Proceedings of LREC 2012*, 2447–2452. İstanbul: ELRA.
http://www.lrec-conf.org/proceedings/lrec2012/pdf/200_Paper.pdf

Válková, L. – Waclawičová, M. – Křen, M. (2012). Balanced data repository of spontaneous spoken Czech. In: *Proceedings of LREC 2012*, 3345–3349. İstanbul: ELRA.
http://www.lrec-conf.org/proceedings/lrec2012/pdf/179_Paper.pdf

Vondřička, P. (2014). Aligning Parallel Texts with InterText. In: *Proceedings of LREC 2014*, 1875–1879. Reykjavík: ELRA.
http://www.lrec-conf.org/proceedings/lrec2014/pdf/285_Paper.pdf