Úvod do korpusové lingvistiky 6

MLUVENÉ KORPUSY
PRAHA – BRNO - OLOMOUC

Korpusy dostupné z ČNK

Část názvu nebo popisu

- bmk 596k
- dialekt v2 – dial 310k
- dialekt v2 – ort 299k
- dialekt v1 – dial 128k
- dialekt v1 – ort 126k
- Jazyky v migraci 294k
- lindsei_cz 135k
- oral v1 6.36M
- oral2013 3.29M
- oral2008 1.35M
- oral2006 1.31M
- orator v2 1.54M
- orator v1 736k
- ortofon v1 1.24M
- ortofon v2 2.56M

- parlcorp 38.6M
- pmk 819k
- schola2010 1.05M
- speeches 249k

Rozsah a synchronnost mluvených korpusů

Korpusy mluveného jazyka (synchronní)

korpus	velikost (počet slov)	lemmatizace	morfologické značky	rok zveřejnění	charakteristika korpusu
Obecné korpusy					
ORATOR (verze 2)	1,2 mil.	✓	✓	2019	referenční korpus monologů s jednoúrovňovou transkripcí
ORTOFON (verze 2)	2,1 mil.	✓	✓	2017	referenční reprezentativní korpus neformální mluvené češtiny s dvouúrovňovou transkripcí (zahrnuje Čechy, Moravu a Slezsko)
ORAL (verze 1)	5,4 mil.	✓	✓	2017	referenční korpus neformální mluvené češtiny (zahrnuje Čechy, Moravu a Slezsko)
ORAL2013	2,8 mil.	✗	✗	2013	referenční reprezentativní korpus neformální mluvené češtiny (zahrnuje Čechy, Moravu a Slezsko)
ORAL2008	1 mil.	✗	✗	2008	referenční sociolingvisticky vyvážený korpus neformální mluvené češtiny (zahrnuje pouze Čechy)
ORAL2006	1 mil.	✗	✗	2006	referenční korpus neformální mluvené češtiny (zahrnuje pouze Čechy)
Specializované korpusy					
BMK	490 tis.	✗	✗	2002	Brněnský mluvený korpus: přepis nahrávek brněnské mluvy z 90. let 20. století
DIALEKT (verze 2)	223 tis.	✓	✓	2017	referenční nářeční korpus s dvouúrovňovou transkripcí
Jazyky v migraci	294 tis.	✓	✓	2022	korpus rozhovorů (vedených v češtině a němčině) s pozdními německými vysídlenci a českými migranty z Československa do Německa o jejich jazykových biografiích
LINDSEI_CZ	120 tis.	✗	✗	2017	žákovský korpus spontánní mluvené angličtiny pokročilých mluvčích s češtinou jako L1
PMK	675 tis.	✗	✗	2001	Pražský mluvený korpus: přepis nahrávek pražské mluvy z 90. let 20. století
SCHOLA2010	790 tis.	✗	✗	2010	korpus vyučovacích hodin
SPEECHES	215 tis.	✓	✓	2015	korpus prezidentských projevů
Parlcorp	38 mil.	✓	✓	2021	korpus projevů v poslanecké sněmovně (1993-2021)

PMK

- **Pražský mluvený korpus** (PMK) je prvním korpusem mluvené češtiny a zachycuje autentickou mluvenou češtinu, hlavně obecnou a tématicky nespecializovanou, resp. neomezovanou, z oblasti Prahy a jejího okolí. Vzhledem k centrálnímu a jedinečnému postavení Prahy tu jazykově dochází k velkému míšení lidí ze všech oblastí ČR a obraz jejího jazyka má tudíž do značné míry celonárodní povahu; z Prahy vychází také nejvýznamnější mediální ovlivnění celé země. Magnetofonové nahrávky (v počtu 304), které jsou plně anonymní a byly postupně přepisovány do počítače, pocházejí z let 1988-1996 a odrážejí tedy jazyk jak konce předchozího společenského období tak začátek nového.

BMK

- **Brněnský mluvený korpus** (BMK) je v rámci ČNK prvním korpusem mluvené češtiny z oblasti Moravy. Zaznamenává autentickou tematicky nespecializovanou mluvu města Brna. BMK je elektronickým přepisem 250 anonymních magnetofonových nahrávek z let 1994-1999 zachycujících 294 mluvčích.
- Značná pestrost brněnské mluvené češtiny odráží složitost sociální struktury velkoměsta, ústřední postavení Brna v rámci Moravy (dochází zde k míšení obyvatel z celého dosud nářečně diferencovaného regionu) a dále teritoriální blízkost k jazykovému území vlastních Čech. V běžné mluvě Brňanů se prolíná zejména středomoravský interdialekt s pronikající obecnou češtinou (s níž se v řadě rysů tradiční dialekt okolí města shoduje), v oblasti slovní zásoby jsou patrny relikty někdejšího soužití brněnské češtiny s německým jazykem a vliv brněnského slangu (hantecu). Mluvený jazyk v Brně reflektuje také celomoravskou tendenci širšího funkčního využití češtiny spisovné.

ORAL2006

- Mluvený korpus **ORAL2006** je v pořadí třetím mluveným korpusem, který je dostupný v rámci projektu Český národní korpus. Zachycuje mluvenou češtinu z celé oblasti českých nářečí v užším slova smyslu. Jedná se o přepis 221 nahrávek z let 2002 - 2006. Všechny nahrávky vznikaly v neformálních situacích, to znamená, že se mluvčí vzájemně znali a měli k sobě přátelský vztah. Celkem bylo nahráno 6 693 minut, tj. asi 111 a půl hodiny, a v jejich rámci zaznamenáno 1 000 798 slov od 754 mluvčích.

ORAL2008

- Korpus **ORAL2008** představuje v rámci projektu Český národní korpus v pořadí již čtvrtý korpus mluvené češtiny. Zachycuje stejně jako ORAL2006 mluvu ve výhradně neformálních situacích. Jde však o první mluvený korpus ÚČNK, který je plně vyvážený v základních sociolingvistických kategoriích mluvčích (pohlaví, věková skupina, výše dosaženého vzdělání a oblast pobytu v dětství). Korpus ORAL2008 vychází ze stejné materiálové základny jako ORAL2006, avšak žádný z přepisů zařazených do korpusu ORAL2008 nebyl použitý v korpusu ORAL2006.

ORAL2008

- Korpus je sestaven z přepisů 297 nahrávek, které byly v letech 2002-2007 pořízeny na různých místech po celém území Čech (tj. ne Česka). Zachycují autentickou mluvenou češtinu v přirozeném prostředí na území tradičně vymezovaném jako oblast českých nářečí v užším smyslu. Vzhledem k postupu nivelizačních procesů jde v projevech nejčastěji o obecnou češtinu a její regionální varianty. Všem nahrávkám je společné to, že byly pořízeny výhradně v neformálních situacích, mluvčí se vzájemně znali a měli k sobě přátelský vztah. Mluvčí nebyli předem informováni o účelu nahrávání, ten jim byl sdělen až po ukončení nahrávání. Všichni následně souhlasili s použitím nahrávky pro potřeby Českého národního korpusu. Nahrávky pro ORAL2008 představují 6 883 minut, tj. necelých 115 hodin, a v jejich rámci byly zaznamenány projevy 995 mluvčích. Celý korpus zahrnuje 1 000 097 slov.

ORAL2013

- Korpus ORAL2013 se skládá z **835 nahrávek** z let **2008–2011** a obsahuje **2 785 189 textových slov**, tj. celkem **3 285 508 pozic**; v sondách vystupuje celkem **2 544 mluvčích**, z toho **1 297 unikátních**. Nahrávky byly pořizovány v Čechách, na Moravě i ve Slezsku, jejich celková délka je **17 471 minut**, tj. téměř 300 hodin.

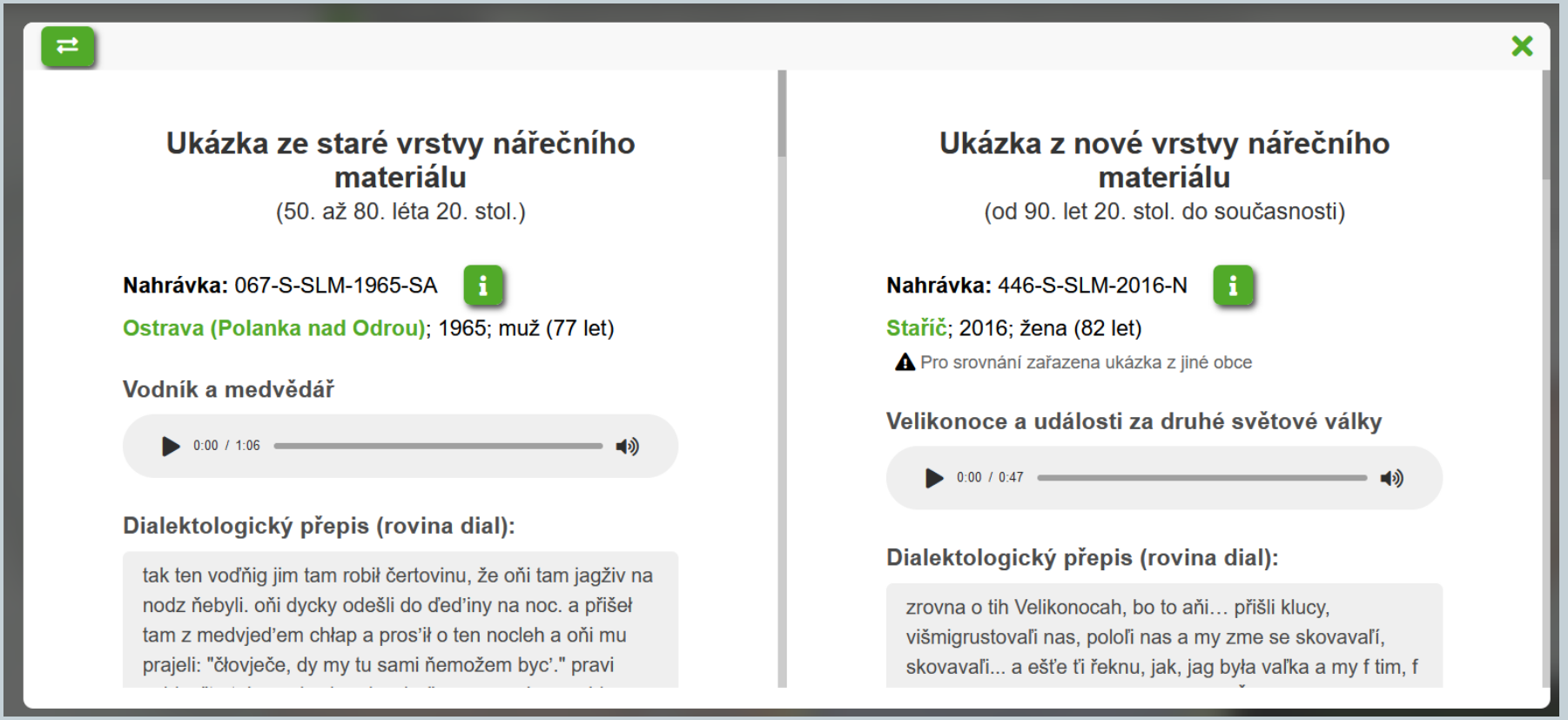
ORTOFON 2017

- ortografická transkripce
- zjednodušená fonetická transkripce
- je možné spustit nahrávku KWIC: https://www.korpus.cz/kontext/view?viewmode=kwic&pagesize=40&attrs=word&attr_vmode=visible-kwic&base_viewattr=word&refs=%3Ddoc.id&q=~SWWc4ekWEA8W&cutoff=0

DIALEKT

- https://wiki.korpus.cz/doku.php/cnk:dialekt
- Nahrávky pro ČJA + další sběr
- 50.-80.léta XX. stol.
- 90. l. XX. stol. – XXI. stol.
- Aplikace MAPKA
- **zvukový záznam** nářečního projevu,
- dialektologický a ortografický **přepis projevu**,
- **rozbor nářečních jevů** obsažených v ukázce,
- **sociolingvistické údaje** o nahrávce a mluvčích

MAPKA


Ukázka ze staré vrstvy nářečního materiálu
(50. až 80. léta 20. stol.)
Nahrávka: 067-S-SLM-1965-SA
Ostrava (Polanka nad Odrou); 1965; muž (77 let)
Vodník a medvědář
0:00 / 1:06
Dialektologický přepis (rovina dial):
tak ten voďňig jim tam robił čertovinu, že oňi tam jagživ na nodz ňebyli. oňi dycky odešli do ďed'iny na noc. a přišeł tam z medvjed'em chłap a pros'ił o ten nocleh a oňi mu prajeli: "človječe, dy my tu sami ňemožem byc'." pravi
Ukázka z nové vrstvy nářečního materiálu
(od 90. let 20. stol. do současnosti)
Nahrávka: 446-S-SLM-2016-N
Stařič; 2016; žena (82 let)
Pro srovnání zařazena ukázka z jiné obce
Velikonoce a události za druhé světové války
0:00 / 0:47
Dialektologický přepis (rovina dial):
zrovna o tih Velikonocah, bo to aňi… přišli klucy, višmigrustovaľi nas, poloľi nas a my zme se skovavaľí, skovavaľi... a ešťe ťi řeknu, jak, jag była vaľka a my f tim, f

DIALEKT


Korpus: dialekt v2 – dial | Dotaz: ešče (186 výskytů) ~ Podrobnosti
Výskytů: 186 | i.p.m.: 599,61 (vztaženo k celému korpusu) | ARF: 69,93 | Výsledek je setříděn
1 / 5
Výběr řádků: základní
0032-C-CEM-1964-SA
0032-C-CEM-1964-SA
0042-M-STM-1966-SA
0042-M-STM-1966-SA
0050-M-STM-1968-SA
0052-M-STM-1968-SA
0052-M-STM-1968-SA
0052-M-STM-1968-SA
0055-M-VYM-1970-SA
0055-M-VYM-1970-SA
0056-M-VYM-1964-SA
0058-M-VYM-1965-SA
0058-M-VYM-1965-SA
0061-M-VYM-1966-SA
Výchozí zobrazení | Promluvy
... • choďil na tu pokrejwačinu , tak já sem nemňela kalčeňi , diš sem nemňela kalčeňi , taki bila bída wo to , tak já sem nanosila za celi leto , sem si uživila pec , co sem spálila , a ešče na zimu .
Aloisie Š. • (nadechnutí)
Aloisie Š. • a topila sem s tim až do ledna , celej leden ešče , jo ?
Vlastimil S. • tos tu pedz živila ?
Aloisie Š. • no , živila tim , tim harapáťim .
Aloisie Š. • bil tenkrát hroznej polom , tadi vítr a
m spálila , a ešče na zimu . + (nadechnutí) + a to
, celej leden ešče , jo ? + tos tu pedz živila ? + no
+ a pag ešče vone vešle a pré : + " poď ešče , pré
+ " poď ešče , pré ten zbetek pré ťe nevemňeňi , za ten
tech kritu , a ešče si vzale chleba a he šlihovice . + (nade
n ve válce a ešče náz zostało šest , štiri ďefčata a dvá kluci .
+ a to ešče nebiło fšecko ňic . + zaz disi zme bili s
tag zaljéčili , ešče pri : + " pudu do teho lesa . " +
, co ďeuáte , ešče potpálíte ! " + " toš co ďeuáme ? šak sr
dajte , ľebo ešče vihoríme . " + (smích více mluvčích naj
sa tam šuo , ešče biuo prítmo . + (nadechnutí) +
múj taťínek ešče , to já uš si tag nepamatuju , ale múj
o aj já doma ešče si ... a randlíki sme m'eli a ti trajfúze a
ťedať prácu , ešče hen na Pústevňe , tam se ďelała nová tá skakač

SCHOLA2010

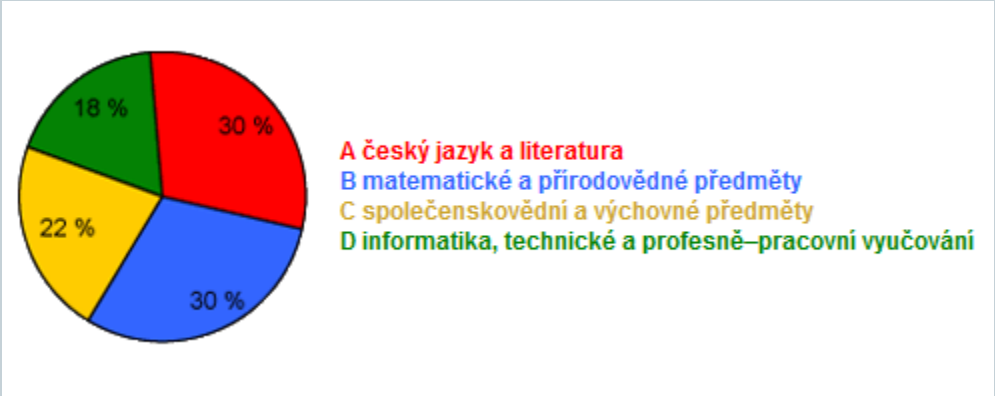
- Řešitelem korpusu SCHOLA2010 je v rámci výzkumného záměru MSM 0021620825 (Jazyk jako lidská činnost, její produkt a faktor) **Ústav českého jazyka a teorie komunikace (ÚČJTK) UK FF**. Jedná se o sociologicky i didakticky jedinečný korpus, protože vychází ze školního prostředí a zaznamenává mluvený jazyk vyučovacích hodin (především standardních vyučovacích hodin s délkou cca 45 min.). Uživatelům se nabízí jazykový materiál, v němž je zachycena mluva učitelů i žáků během vyučování. Zatím je to jediný veřejně přístupný korpus tohoto typu. Uvedený korpus se od ostatních mluvených korpusů zveřejněných v Českém národním korpusu (ČNK) liší také tím, že obsahuje mluvu dětí a mládeže.

SCHOLA2010

- Korpus SCHOLA2010 tvoří **204 přepisů nahrávek vyučovacích hodin**, pořízených v letech **2005–2008**. Sondy pocházejí z různých míst České republiky, viz oddíl Statistiky ke korpusu Schola2010. 131 nahrávek bylo nahráno ve středočeské nářeční oblasti, 57 nahrávek ve východomoravské nářeční oblasti (vymezení nářečních oblastí se opírá o pojetí Běličovo, *Nástin české dialektologie*, 1972, a o členění nářečních oblastí v Českém jazykovém atlasu, 1992–2005, viz mapa nářečních oblastí podle ČJA), jde tedy i o teritoriálně různorodý jazykový materiál.

Korpus vyučovacích hodin SCHOLA2010

- Zastoupení dle vyučovacích předmětů



PARLCORP

- **Korpus českých parlamentní projevů**

Korpus: parlcorp | **Dotaz:** důchod (7 161 výskytů) ~ Podrobnosti

Výskytů: **7 161** | i.p.m.: **185,56** (vztaženo k celému korpusu) | ARF: **1 313,26** | Výsledek je setříděn

1 / 180

Výběr řádků: základní

doc#1	ITRL . Tento ekonomický ukazatel vlastně ukazuje míru celkových odvodů	důchodů	obyvatel do různých odvodů daňových , sociálních , pojistných ,
doc#1	mateřské dovolené , rodičovského příspěvku , nemocenské , ale i	důchody	, řešili jsme to . To jsou ti ? Těch
doc#2	nelze posuzovat obě dvě situace , to je odchod do	důchodu	a způsobilost k řízení stejným měřítkem . (Potlesk z
doc#2	spojit , jak navrhuji , právě s řádným odchodem do	důchodu	. Myslím si , že na tom nic není .
doc#2	tím nesouhlasím - aby se neustále zvyšoval věk odchodu do	důchodu	. To je naprosto nesouměřitelné a já si myslím ,
doc#13	Slovenské republice , než kdyby tyto doby byly zhodnoceny v	důchodech	českých . Z tohoto důvodu se dotčené osoby dlouhodobě domáhaly
doc#13	podpory poslanců Poslanecké sněmovny Parlamentu České republiky , přiznání českého	důchodu	za tyto doby . Proto se Poslanecká sněmovna i Senát

Korpus prezidentských projevů

- speeches
  https://wiki.korpus.cz/doku.php/cnk:speeches
- *written to be spoken*

Zásady přepisu

- pravopis a přepis x fonetický přepis
- interpunkce „pauzová“
- pravopis i-y
- pravopis ě
- délka vokálů
- asimilace znělosti
- zdvojené souhlásky
- artikulační asimilace
- cizí slova a propria
- přitákání, odmítnutí, smích, komentáře, nesrozumitelné úseky, překryvy

Sociolingvistické značkování

- pohlaví
- věk
- vzdělání
- teritoriální zařazení

Ochrana poskytovatelů

- prohlášení dovolující uveřejnění nahrávky za stanovených podmínek
- vynechání veškerých údajů, přes něž by bylo možné „vysledovat“ hovořící (jména, adresy, tel. čísla, …)

Olomoucký mluvený korpus

- Projekt dr. P. Pořízky UPOL
- foneticky přepsané texty
- http://korpling.webnode.cz/olomoucky-mluveny-korpus/

Publikace

Čeština v mluveném korpusu

Marie Kopřivová a Martina Waclawičová

Ke značkování mluvených korpusů

- HLAVÁČKOVÁ, Dana a Klára OSOLSOBĚ. Morfologické značkování mluvených korpusů, zkušenosti a otevřené otázky. Kopřivová, Marie, Waclawičová, Martina. In *Čeština v mluveném korpusu*. 1. vyd. Praha: Nakladatelství Lidové noviny/ Ústav Českého národního korpusu, 2008. s. 105-114, 10 s. ISBN 978-80-7106-982-9.
- https://wiki.korpus.cz/doku.php/cnk:lemtag_mluv
- Kopřivová, M. - Lukeš, D. - Komrsková, Z. - Poukarová, P. (2017): Korpus ORAL: sestavení, lemmatizace a morfologické značkování. In *Korpus - Gramatika - Axiologie*, 15, 47-67.
- Lukeš. D. - Klimešová, P. - Komrsková, Z. - Kopřivová, M. (2015) : Experimental Tagging of the ORAL Series Corpora: Insights on Using a Stochastic Tagger. In: TSD 2015, Ed. P. Král a V. Matoušek. Springer international Publishing, 342-350.

Publikace

Morfologie mluvené češtiny: Frekvenční analýza

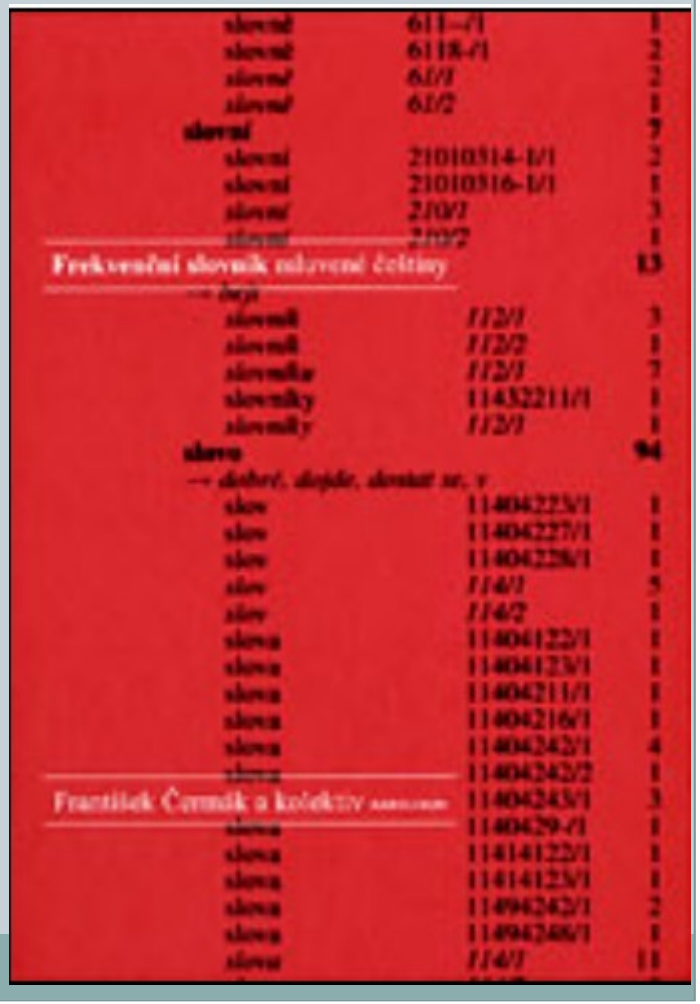
Jitka Šonková

Tento svazek podává první soustavnou charakteristiku skloňování a časování v mluvené češtině. Studie vychází z kvantitativní analýzy Pražského mluveného korpusu, tvořeného přepisy více než 304 nahrávek z Prahy a okolí, a zaměřuje se především na konkurenci spisovných a nespisovných tvarů v běžné komunikaci českých mluvčích.

Šonková, J.: Morfologie mluvené češtiny: Frekvenční analýza. Nakladatelství Lidové noviny, Praha 2008.
ISBN 978-80-7106-956-0

Publikace

- Čermák, F. (ed.): Frekvenční slovník mluvené češtiny. Karolinum, Praha 2007. ISBN 978-80-246-1425-0



Příklady otázek v testu

- Který byl první mluvený korpus zveřejnění ÚČNK?
- Která další pracoviště se podílí na budování korpusů mluveného jazyka zveřejněných ÚČNK?
- V jaké podobě jsou uložena data mluvených korpusů (typy transkripce, přístup k nahrávkám)?
- Které publikace vznikly na základě mluvených korpusů češtiny?
- Které z mluvených korpusů jsou lemmatizovanéa morfologicky označkované?