

Úvod do korpusové lingvistiky 6



MLUVENÉ KORPUSY
PRAHA – BRNO - OLMOUC

Korpusy dostupné z ČNK



- Synchronní mluvené korpusy
- Řada ORAL
- Specializované (mluvčí Prahy a Brna)
- Schola

- ▼ Synchronní mluvené korpusy
 - ▼ řada ORAL
 - [oral2008](#)
 - [oral2006](#)
 - ▼ specializované
 - [bmk](#)
 - [pmk](#)
 - [schola2010](#)

Rozsah a synchronnost mluvených korpusů



Korpusy mluveného jazyka (synchronní)

korpus	velikost (počet slov)	lemmatizace	morfologické značky	rok zveřejnění	charakteristika korpusu
<u>ORAL2013</u>	2,79 mil	NE	NE	2013	reprezentativní korpus neformální mluvené češtiny
<u>ORAL2008</u>	1 mil	NE	NE	2008	sociolingvisticky vyvážený korpus neformální mluvené češtiny
<u>ORAL2006</u>	1 mil.	NE	NE	2006	korpus neformální mluvené češtiny
<u>SCHOLA2010</u>	790 000	NE	NE	2010	korpus vyučovacích hodin
<u>PMK</u>	675 000	NE	NE	2001	Pražský mluvený korpus: přepis nahrávek pražské mluvy z 90. let 20. století
<u>BMK</u>	490 000	NE	NE	2002	Brněnský mluvený korpus: přepis nahrávek brněnské mluvy z 90. let 20. století

PMK



- **Pražský mluvený korpus (PMK)** je prvním korpusem mluvené češtiny a zachycuje autentickou mluvenou češtinu, hlavně obecnou a tématicky nesespecializovanou, 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 nesespecializovanou 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 patrné 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 vymezeném 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.

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.

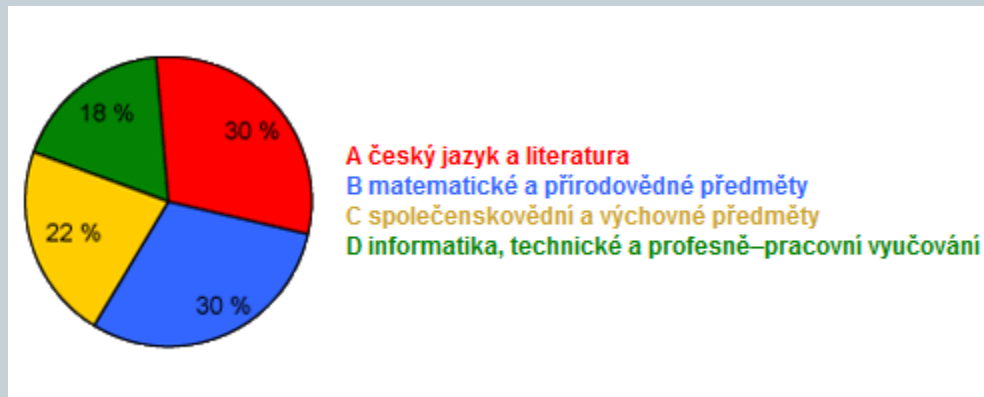
SCHOLA2010



- Korpus SCHOLA2010 tvoří **204 přepisů nahrávek vyučovacích hodin**, pořizených 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



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

Sociolinguvistické 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.

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 přepisem 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

