ČASOPIS PRO MODERNÍ FILOLOGII 97, 2015, Č. 1, S. 42–50

Korpus spontánní mluvené češtiny ORAL2013[1]

Lucie Benešová, Michal Křen, Martina Waclawičová (Praha)

THE CORPUS OF SPONTANEOUS SPOKEN CZECH ORAL2013

The paper presents a corpus of spontaneous spoken Czech called ORAL2013, its design principles and practical solutions adopted during the data collection. The corpus is designed to represent contemporary spontaneous spoken language used in informal, real-life situations across the whole of the Czech Republic. The corpus consists of audio recordings and their transcriptions aligned with time stamps; it features manual annotation and broad regional coverage with a large variety of speakers. ORAL2013 contains 835 recordings from the period 2008 to 2011 made with 2,544 speakers (of whom 1,297 speakers are unique); the total length of the audio tracks is almost 300 hours and the total size of the transcriptions exceeds 3.28 million tokens. ORAL2013 is made publicly available by the Czech National Corpus at http://www.korpus.cz/.

KEYWORDS

language corpus, corpus design, spontaneous spoken language, Czech, transcription

KLÍČOVÁ SLOVA

jazykový korpus, složení korpusu, spontánní mluvený jazyk, čeština, transkripce

1. ÚVOD

Mluvený jazyk je primárním prostředkem naší každodenní komunikace, a mimo jiné proto si získává stále větší pozornost lingvistů. Stěžejním tématem je v této souvislosti dichotomie mluveného a psaného jazyka, které byla věnována řada diskuzí a prací (např. Vachek, 1989; Tannen, 1982; Chafe, 1980; Chafe — Danielewicz, 1987 aj.). Odlišnost psaného jazyka od jazyka mluveného (především spontánního) se projevuje na mnoha jazykových úrovních: „The terms 'spoken language' and 'written language' do not refer merely to different mediums but relate to partially different systems of morphology, syntax, vocabulary, and the organization of texts" (Miller — Weinert, 1998, s. 5).

Pod pojmem spontánní mluvený jazyk rozumíme interakci, která vzniká při bezprostředním kontaktu s komunikačním partnerem, v neformálním a neveřejném prostředí, v reálném čase, bez přípravy a bez možnosti do výsledného produktu jakkoli zpětně zasahovat a která je limitována kapacitou krátkodobé paměti mluvčího i posluchače a je silně vázána na konkrétní situační kontext (Miller — Weinert, 1998; Čmejrková — Hoffmannová, 2011). Tyto „produkční podmínky" jsou jednou z pří-

1 Tento článek vznikl při realizaci projektu Český národní korpus (LM2011023) financovaného Ministerstvem školství, mládeže a tělovýchovy v rámci aktivity Projekty velkých infrastruktur pro VaVaI.

čin toho, že spontánní konverzace se obecně vyznačuje chudší slovní zásobou než produkce psaná, syntax spontánních mluvených komunikátů je fragmentarizovaná, celkově méně komplexní (viz např. Hoffmannová, 2012) a významnou roli hraje — hlavně díky kontextu — deixe (Hirschová, 2013).

Chceme-li spontánní mluvený jazyk zkoumat, je nutné mít k dispozici spolehlivá a reprezentativní data. Pojem reprezentativnosti jazykového korpusu je ovšem definován různě a také míru, v níž korpus jazyku odpovídá, nelze objektivně měřit. Přesto je reprezentativnost korpusu potřebná, protože umožňuje zobecnění výsledků výzkumu založeného na korpusu jako vzorku jazyka pro jazyk jako celek. Korpus ORAL2013, který v tomto článku detailněji představíme, si klade nelehký cíl být reprezentativním korpusem spontánní mluvené češtiny a umožnit tak studium této její dosud málo prozkoumané variety.

2. KORPUSY ŘADY *ORAL*

Korpus ORAL2013 (Válková et al., 2012) je dalším z korpusů spontánní mluvené češtiny řady ORAL. Předcházely mu korpusy ORAL2006 (Kopřivová — Waclawičová, 2006) a ORAL2008 (Waclawičová et al., 2009) čítající každý 1 milion textových slov. Oba byly vybudovány na společné materiálové základně — aniž by ale měly společný průnik — složené ze sond pořízených na různých místech na území Čech.

Oba tyto korpusy slouží jako hodnotné a hojně využívané zdroje přispívající k popisu neformální mluvené češtiny. Jejich výpovědní hodnota je však omezena tím, že pokrývají pouze oblast českých nářečí v užším smyslu, tj. Čechy bez Moravy a Slezska; také přepis nahrávek je v zásadě ortografický a neumožňuje věrné zachycení řady fonetických jevů. V neposlední řadě práci s oběma korpusy ztěžuje fakt, že přepis není nijak propojen se zvukem.

ORAL2013 se proto od svých předchůdců liší především v manuálním propojení textového přepisu se zvukovou stopou pomocí programu *Transcriber* (Geoffrois et al., 2000) a v tom, že jeho materiálová základna pokrývá území celé České republiky, tedy všechny oblasti Čech, Moravy a Slezska. To s sebou zároveň přineslo nutnost rozpracovat přepisovací pravidla tak, aby byl přepis schopen zachytit i regionální zvláštnosti nově zpracovávaných území a konečně došlo k přechodu od tradiční syntaktické interpunkce k interpunkci pauzové.

3. ZPŮSOB SBĚRU DAT

Nutným předpokladem pro vytvoření reprezentativního korpusu spontánního mluveného jazyka je dostatečné množství jazykového materiálu zachycujícího velký počet různorodých mluvčích s co nejúplnějším regionálním pokrytím (Gibbon et al., 1998). Sběr materiálu pro korpus ORAL2013 proto probíhal kromě Českého národního korpusu (ČNK) také na pracovištích Masarykovy univerzity, Univerzity Hradec Králové, Univerzity Palackého v Olomouci a Západočeské univerzity v Plzni; ve spolupráci s nimi vznikla síť spolupracovníků, jejichž celkový počet dosáhl téměř dvou set.

Sběr dat pro ORAL2013 byl z organizačních důvodů rozdělen hierarchicky do následujících tří úrovní:

1. *Hlavní koordinátoři* projektu v ČNK byli zodpovědní za celkovou organizaci, administrativu a technickou podporu, včetně centrálního úložiště, standardizace a kvalitativních kontrol jazykového materiálu. Součástí hlavní koordinace bylo také pořádání školení, tvorba a aktualizace manuálů apod.
2. *Místní koordinátoři* na jednotlivých univerzitách byli zodpovědní za svoji oblast sběru, zejména za složení a celkovou kvalitu dat; v jejich pravomoci byl proto výběr a zacvičení konkrétních spolupracovníků (editorů) spolu s kontrolami jejich práce.
3. *Editoři* byli většinou studenti, kteří nahrávky pořizovali a ve většině případů je také přepisovali.

Pro práci na projektu bylo pro jeho široký záběr důležité odpovídající technické zázemí. Proto vzniklo databázové rozhraní *Mluvka* (Křen — Waclawičová, 2011), jehož prostřednictvím byl postupně ukládán veškerý shromažďovaný materiál, a to ve formě jednotlivých sond; každá sonda sestávala vždy ze zvukové nahrávky, jejího přepisu a souvisejících metadat. Každý spolupracovník měl v rozhraní svůj účet s přístupovými právy omezenými vždy jen na sondy, s nimiž potřeboval pracovat.

Protože sběr materiálu pro ORAL2013 probíhal výhradně prostřednictvím databázového rozhraní *Mluvka*, byly jeho součástí také všechny potřebné kontroly, zejména kontrola formální správnosti přepisů. Ta zahrnovala prakticky všechny rysy zmiňované v přepisovacích pravidlech, které bylo možné kontrolovat automaticky, například správnou segmentaci anonymizačních zkratek (kvůli následné automatické anonymizaci odpovídající části zvukového souboru), používání náležitých značek pro hezitační a responzní zvuky, konzistentní přepis některých problematických jevů (např. nespisovný kondicionál 1. os. pl. zapisovaný výhradně bez mezery jako *bysme*) apod.

Typický postup zpracování sondy začínal jejím vložením editorem do rozhraní *Mluvka* a pokračoval jejím předáváním v rámci tohoto rozhraní přes místního koordinátora ke koordinátorovi hlavnímu do ČNK. Celou sondu bylo možné předat dál pouze tehdy, pokud byla vyplněna kompletní metadata, nahrávka byla ve správném formátu a přepis prošel všemi formálními kontrolami. Kontrola ze strany koordinátorů se tak mohla soustředit na věrnost přepisu nahrávce a korektní zachycení relevantních jazykových jevů v něm; v případě zjištění jakýchkoli nedostatků mohli koordinátoři sondu vracet zpět k přepracování. Všechny sondy tedy po svém vzniku prošly kontrolami nejméně na dvou úrovních, jejichž cílem bylo zajistit maximální správnost a konzistenci přepisu.

Metadata uložená v databázovém rozhraní *Mluvka* je možné rozdělit do dvou skupin:

1. údaje o sondě jako celku, a to jak rázu spíše technického (délka nahrávky, měsíc, rok a místo nahrávání apod.), tak i odborného, popisující zejména konkrétní typ komunikační situace (návštěva, restaurace, oslava, výlet apod.), téma rozhovoru, počet mluvčích v sondě a jejich vzájemný vztah. Vybrané údaje této skupiny jsou v korpusu ORAL2013 dostupné jako hodnoty příslušných atributů struktury <*doc*>.

2. údaje o mluvčích: pohlaví, věk, úroveň dosaženého vzdělání, současné zaměstnání, místo narození, oblast narození, oblast pobytu v dětství a oblast současného pobytu. Jednotlivé oblasti jsou vymezeny (stejně jako u stávajících korpusů řady ORAL) na základě tradičního nářečního členění podle Jaromíra Běliče a dělení používaného v Českém jazykovém atlase (Bělič, 1972; Balhar, 1992–2011). Vybrané údaje této skupiny jsou v korpusu ORAL2013 dostupné jako hodnoty příslušných atributů struktury <*sp*>.

Rozhraní *Mluvka* kromě základních funkcí, jako jsou prohledávání a úpravy metadat, ukládání nahrávky nebo již zmíněné formální kontroly jejího přepisu, poskytovalo také funkce rozšířené, které se v průběhu práce na korpusu ukázaly jako velice důležité. Šlo především o zjišťování počtu slov vyslovených jednotlivými mluvčími v každé sondě; tato informace byla nezbytná pro vyvažování shromážděného materiálu. Počty slov jednotlivých mluvčích byly totiž propojeny s jejich sociolingvistickými kategoriemi a počty slov v těchto kategoriích pak mohly být zobrazeny v souhrnné podobě pro libovolně definované množiny sond, což umožňovalo průběžnou kontrolu vyváženosti dat na různých úrovních.

Další důležitou rozšířenou funkcí byla poloautomatická detekce mluvčích vystupujících ve více nahrávkách, kterou bylo nutné implementovat proto, že do databáze *Mluvka* nebyly o mluvčích záměrně ukládány žádné osobní údaje, jež by umožnily jejich identifikaci. Tato detekce průběžně vyhledávala potenciálně shodné dvojice mluvčích a nabízela je hlavním koordinátorům k posouzení. Takto vzniklé proznačení shodných mluvčích sloužilo k omezení maximálního počtu slov každého konkrétního mluvčího v celém korpusu již při sběru dat, a to na 15 000 textových slov; v současné době je přístupné uživatelům korpusu ORAL2013 ve formě náhodně vygenerované „přezdívky“ (strukturní atribut *oznacenishody*), která je pro daného mluvčího v celém korpusu stejná.

4. PŘEPISOVACÍ PRAVIDLA

Způsob transkripce korpusu ORAL2013 byl koncipován s ohledem na tři důležité faktory: na zachování kontinuity se stávajícími korpusy řady ORAL (tj. ORAL2006 a ORAL2008), na potřebu zapojit velký počet přepisujících a konečně na technické možnosti zobrazení a priority výzkumu. Zdůrazněme proto, že korpus ORAL2013 není určen primárně pro fonetický nebo dialektologický výzkum, ale že je navržen tak, aby umožňoval zkoumat především morfologii, syntax/syntagmatiku, lexikum a pragmatiku mluveného jazyka, což je podpořeno relativně velkým objemem získaných dat; vhodný je také pro výzkum struktury spontánního mluveného diskurzu.

Přepis je jednoúrovňový, v zásadě ortografický, ručně propojený se zvukovou stopou. Důležité metatextové informace vztahující se ke komunikační situaci (smích, pláč, kýchnutí aj.) jsou uvedeny jako komentář v kulatých závorkách.

Repliky mluvčích jsou členěny na segmenty představující sémanticky, prozodicky nebo syntagmaticky ucelenou sekvenci v průměru o 5–10 slovech. Maximální délka jednoho segmentu byla omezena na 15 slov. Segmenty jsou ve webovém rozhraní *Kon-*

Text vizuálně označeny uzavřením do hranatých závorek, které umožňují příslušný segment také přehrát a poslechnout si tak jeho skutečnou realizaci; zkratkou pro sekvenční přehrání dvou sousedních segmentů je znaménko plus.

Jednotliví mluvčí jsou v přepisech označeni čísly, nulou je vždy odlišen mluvčí, který nahrávku pořizoval, a tudíž o nahrávání nutně věděl. Označeny jsou také simultánní úseky, v nichž mluví dva mluvčí současně. V korpusu je tato informace uživatelům dostupná jako hodnota strukturního atributu *prekryv*; na první pohled patrným indikátorem překryvu je chybějící znaménko plus, které se objevuje pouze mezi nepřekrývajícími se segmenty.

Kvůli respektování ochrany osobních údajů jsou veškerá příjmení a telefonní čísla v přepisech kódována anonymizačními zkratkami; kódování ostatních vlastních jmen, jako jsou např. přezdívky, rodná jména, názvy firem, případně jiné citlivé údaje, bylo ponecháno na vůli a přání přepisujících nebo samotných mluvčích. Anonymizovány byly samozřejmě i odpovídající úseky ve zvukových souborech.

Tradiční syntaktickou interpunkci nahradila interpunkce pauzová, která je pro přepis spontánní mluvené konverzace mnohem vhodnější. Rozlišují se celkem tři typy pauz: krátká pauza (v přepisech označená jednou tečkou), delší pauza (označená dvěma tečkami) a odmlčení (označené jako poznámka v kulatých závorkách). Délku pauz přepisující zaznamenávali podle individuálního tempa řeči každého z mluvčích.

Vlastní přepisovací pravidla vycházejí z přepisovacích pravidel pro korpusy ORAL2006 a ORAL2008, která byla pouze rozšířena o popis toho, jakým způsobem zachytit vybrané specifické jazykové jevy vyskytující se na území Moravy a Slezska. Odlišná výslovnost je v přepisech reflektována v určitých pravidelných, přesně popsaných a stanovených případech, jako je např. zjednodušování souhláskových skupin (*chlapík kerýmu mohlo bejt tak; já ho vemu; esi by nebylo lepší; má pučenej ten traktor*), odlišná délka samohlásek (*těstoviny z makem; zatim platim furt stejně; s kamošem jednim; no bóže; von si cucá svoje pívo*) nebo jiné výslovnostní či regionální variace (*na vokraj ňákej; to prodaj; do štyrycítky asi no; ani houno čoveče; vajca sou drahý; oni sú z něho jako na mrtvicu; jak to vodevřeš tu flašu; vlastně má osnást . sednást roků; záda z toho bolijou; oni to akorát diagnostikujú a nic jiného s tím neudělajú; sem se aj po tem začal smít už; brzo sa žénil no ten tata; voni majó tři děcka; pěkné faň léto; se mi tady podepište kucí; kúsek masa; ten měl čepicu velikú*).

Hvězdičkou jsou v přepisech označeny tyto tři jevy: nedokončená slova (*přijede se k mos* projede se Dubí*), nevyslovený začátek slova (**ce neby* nebyly moc drahý*) a tzv. příklonné *–s* zastupující tvar pomocného slovesa *být* ve 2. os. sg. min. č. (*teď *s mě vzbudil; v pátek *s tam byl; viděla *s ho tam určitě; *s to nechal tak špinavý*).

V přepisech jsou důsledně zaznamenány rysy typické pro nepřipravenou spontánní konverzaci jako např. falešné starty (*mně ňák jak to . mně pak třeští skoro hlava*), fragmentárnost, myšlenková a formulační roztříštěnost (*vzadu to měla jako roztřepaný víš takovej ten účes ona je tmavovláska*), významová neurčitost (*já nejsem ňákej takovej . eee . prtože hele divej se já sem .. takovej už jako*), rektifikace a opravy (*v ten . v tu středu .. v ten čtvrtek*), opakování (*eště nám tam . eště nám tam na štyry pomela zůstalo; to bylo to bylo krásný to bylo*), doplňování a dodávání informací (*a tak bych se někam.. klidně aji zajel podívat jako dyby třeba nejeli do Londýna a jeli by někam . eee .. do Španělska nebo do toho Nizozemska prostě nebo někam tam kde sem eště nebyl . tak bych jel*), užívání prázdných, redundantních výrazů (*pořád takový jako miloučký takový jako jo; a takže prostě no mmm*

no .. prostě a to . jo? to je takový .. taková.. pocta pro toho člověka), hezitace (*no to jo ale eee von to suší v garáži*), kontextové elipsy (*já sem . ještě nebyl .. dycky do jiný*) atp. Nedokončené, přerušené nebo nesrozumitelné úseky jsou zachyceny speciálními značkami.

Odkaz na kompletní znění přepisovacích pravidel je součástí popisu korpusu ORAL2013 na internetové wiki ČNK.[2]

5. SLOŽENÍ KORPUSU

Korpus ORAL2013 je koncipován jako reprezentativní vzorek spontánní mluvené češtiny používané v neformálních komunikačních situacích. Právě tato varieta mluveného jazyka je považována za prototypickou, nejvíce odlišnou od jazyka psaného (Čermák, 2009).

Faktory ovlivňující charakter mluveného jazyka můžeme pro účely sestavení korpusu rozdělit do dvou hlavních skupin. Do první skupiny řadíme faktory převážně situační, které jsou důležité pro zajištění výše zmíněné prototypické neformálnosti, a které tedy tvoří kritéria pro získávání nahrávek a jejich zařazení do korpusu. Klíčovou roli tady hraje neformálnost komunikační situace (Labov, 1972) a její neveřejný a neoficiální charakter; dalšími faktory jsou zejména soukromé prostředí, dialogičnost promluv, fyzická přítomnost mluvčích[3] a jejich vzájemný blízký vztah, nepřipravenost a spontánnost. Všechny nahrávky zařazené do korpusu ORAL2013 tato kritéria splňují, což zajišťuje maximální prototypičnost zachyceného jazyka. Nejčastějšími komunikačními situacemi v korpusu jsou proto návštěva, rozhovor doma, v restauraci, při společné činnosti apod.

Do druhé skupiny řadíme faktory sociolingvistické, které ovlivňují charakter mluveného jazyka v menší míře. Jde především o pohlaví, věk, vzdělání a oblast pobytu v dětství. Tyto čtyři faktory jsme také vybrali jako kritéria pro průběžné vyvažování dat, zatímco ostatní sociolingvistické údaje (např. místo nahrávky, počet mluvčích v sondě, jejich zaměstnání, místa narození atd.) jsou pouze zaznamenány ve formě metadat o sondě a konkrétních mluvčích. Pro zachování kompatibility se staršími korpusy řady ORAL jsme pro průběžné vyvažování sběru dat použili v případě věku a vzdělání pouze zjednodušené binární hodnoty: mladší (18–34 let) / starší (35 a více let) věk a nižší/vyšší vzdělání (za vyšší se považuje VŠ vzdělání i jen započaté); 50% podíl každé z těchto binárních hodnot se tedy považuje za ideální zastoupení dané kategorie.[4]

Takto stanovený cíl průběžného vyvažování korpusu vychází ze způsobu vyvážení korpusu ORAL2008; jde pochopitelně o čísla, která skladbu populace odrážejí pouze přibližně. Na druhou stranu snaha o přesné zohlednění skladby populace by — přes praktická úskalí, která by to přinášelo, zejména v případě oblasti pobytu v dětství — nemusela nutně vést k hodnotnějšímu korpusu. ORAL2013 záměrně není koncipován jako vyvážený v tom smyslu, že by jeho složení mělo ve sledovaných kritériích přesně

2 https://wiki.korpus.cz/doku.php/cnk:oral2013

3 Záměrně jsme nesbírali data z telefonních rozhovorů, z komunikace přes Skype a z jiných podobných situací.

4 Korpus byl vyvažován pouze vzhledem k jednotlivým kategoriím, ne jejich kombinacím.

odrážet skladbu populace. Naším cílem byla „pouze" reprezentativnost ve smyslu *„texts as products"* (Biber, 1993, s. 245), tedy sestavení korpusu jako vzorku dostatečně pokrývajícího variabilitu spontánního mluveného jazyka svým širokým záběrem a maximální rozmanitostí mluvčích. Právě průběžné vyvažování složení korpusu podle těchto přibližných čísel významně napomáhá této rozmanitosti dosáhnout.

Při konečné přípravě dat jsme tedy na rozdíl od korpusu ORAL2008 nepřistoupili k plnému vyvážení korpusu, a to také proto, že by to znamenalo zbavování se cenného materiálu v situaci, kdy je korpus už dostatečně reprezentativní, zatímco jeho hypotetická „ideální" vyváženost by byla jednak sporná (zvláště její regionální faktor) a jednak by nebyla ani potřebná, protože webové rozhraní *KonText* umožňuje práci s relativními, a tedy srovnatelnými frekvencemi.

Přesné složení korpusu v základních sociolingvistických kategoriích je uvedeno v následujících tabulkách:

pohlaví	**ženy**	**muži**
	1 359 761	1 425 428
věk	**mladší**	**starší**
	1 458 386	1 326 803
vzdělání	**nižší**	**vyšší**
	1 515 732	1 269 457

TABULKA 1: Počet textových slov ve vybraných sociolingvistických kategoriích.

oblast pobytu v dětství	**počet textových slov**
středočeská	570 283
severovýchodočeská	353 486
jihozápadočeská	315 716
české pohraničí	191 553
česko-moravská	83 478
středomoravská	503 391
východomoravská	359 249
slezská	317 087
moravské pohraničí	90 946

TABULKA 2: Počet textových slov podle převažující oblasti pobytu v dětství.

Korpus ORAL2013 se skládá z celkem 835 nahrávek z let 2008–2011 a obsahuje 2 785 189 textových slov, tj. 3 285 508 korpusových pozic (tokens včetně interpunkce); 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. 291 hodin. Všechny nahrávky jsou ve formátu 16-bit PCM WAV, mono, 16 kHz.

Naším primárním cílem bylo zachování maximální autenticity mluvčích a jejich projevů, která je ovšem podmíněna přirozeným prostředím. Mluvčí proto většinou nebyli o nahrávání informováni předem, ale až po jeho skončení, kdy také dávali sou-

hlas s poskytnutím nahrávky a jejího přepisu do ČNK. Je tedy zřejmé, že ačkoli jsme dbali na kvalitu zvukových nahrávek, nebylo za těchto podmínek často možné vyhnout se ruchům a šumům, které jsou pro podobné situace typické.

6. ZÁVĚR

V článku jsme představili korpus spontánní mluvené češtiny ORAL2013, jeho koncepci, složení, zpracování a způsob sběru dat. Korpus je pečlivě manuálně anotován a pomocí přepisů propojených se zvukovou stopou zachycuje projevy velkého množství různých mluvčích ve výhradně neformálních komunikačních situacích na území celé České republiky. Jeho povaha, velikost a regionální pokrytí jej řadí k předním korpusům svého druhu na světě. ORAL2013 byl zpřístupněn v prosinci 2013 všem registrovaným uživatelům ČNK na adrese http://www.korpus.cz/. Pro vědecké, pedagogické a jiné nekomerční účely je možné získat také celý korpus s kompletními přepisy a anonymizovanými nahrávkami, což umožňuje jeho využití dalšími způsoby a softwarovými nástroji.

Závěrem bychom chtěli poukázat také na dva nedostatky korpusů řady ORAL. Jde jednak o způsob přepisu, který byl z praktických důvodů pouze jednoúrovňový, a jednak o absenci lemmatizace a morfologického značkování, omezující možnosti práce s tímto cenným materiálem. V blízké budoucnosti proto počítáme se zveřejněním korpusu ORAL, který bude spojením všech korpusů této řady a který bude navíc opatřen lemmatizací a morfologickým značkováním. Kromě toho kolegové v ÚČNK usilovně pracují na korpusu nové generace ORTOFON se dvěma úrovněmi anotace, ortografickou a fonetickou (Kopřivová et al., 2014), který bude moderním pokračováním řady mluvených korpusů ČNK.

7. PODĚKOVÁNÍ

Děkujeme všem, kteří se podíleli na pořizování nahrávek, jejich přepisu a následných úpravách, především studentům Filozofické fakulty Univerzity Karlovy v Praze. Sběru materiálu se pod vedením svých pedagogů účastnila také řada studentů Masarykovy univerzity, Univerzity Hradec Králové, Univerzity Palackého v Olomouci a Západočeské univerzity v Plzni. Zvláštní poděkování za skvělou spolupráci patří Haně Voralové.

LITERATURA

Balhar, J. (ed.) (1992–2011): *Český jazykový atlas*. Praha: Academia.

Bělič, J. (1972): *Nástin české dialektologie*. Praha: SPN.

Biber, D. (1993): Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8/4, s. 243–257.

Čermák, F. (2009): Spoken corpora design. Their constitutive parameters. *International Journal of Corpus Linguistics*, 14/1, s. 113–123.

Čmejrková, S. — Hoffmannová, J. (2011): *Mluvená čeština: hledání funkčního rozpětí*. Praha: Academia.

Geoffrois, E. — Barras, C. — Bird, S. — Wu, Z.

(2000): Transcribing with annotation graphs. In: *Proceedings of LREC2000*, s. 1517–1521.

Gibbon, D. — Moore, R. — Winski, R. (eds.) (1998): *Spoken language system and corpus design*. Berlin: Mouton de Gruyter.

Hirschová, M. (2013): *Pragmatika v češtině*. Praha: Karolinum.

Hoffmannová, J. (2012): Syntaktická stylistika mluvených projevů. In: S. Čmejrková — J. Hoffmannová — J. Klímová (eds.), *Čeština v pohledu synchronním a diachronním: Stoleté kořeny Ústavu pro jazyk český*. Praha: Karolinum, s. 707–713.

Chafe, W. L. (1980): The deployment of consciousness in the production of a narrative. In: W. L. Chafe (ed.), *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*. Norwood: Ablex, s. 9–50.

Chafe, W. L. — Danielewicz, J. (1987): Properties of written and spoken language. In: R. Horowitz — S. J. Samuels (eds.), *Comprehending Oral and Written Language*. New York: Academic Press, s. 83–113.

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 LREC2014*, s. 376–382.

Kopřivová, M. — Waclawičová, M. (2006): Representativeness of Spoken Corpora on the Example of the New Spoken Corpora of the Czech Language. In: *Труды международной конференции «Корпусная лингвистика — 2006»*. Санкт-Петербург: Издательство СПбГУ, s. 174–181.

Křen, M. — Waclawičová, M. (2011): Database framework for a distributed spoken data collection project. In: S. Goźdź-Roszkowski (ed.), *Explorations across Languages and Corpora*. Frankfurt am Main: Peter Lang, s. 83–93.

Labov, W. (1972): *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Miller, J. — Weinert, R. (1998): *Spontaneous spoken language. Syntax and discourse*. Oxford: Clarendon Press.

Tannen, D. (1982): Oral and literate strategies in spoken and written language. *Language*, 58, s. 1–21.

Vachek, J. (1989): *Written Language Revisited*. Amsterdam / Philadelphia: John Benjamins.

Válková, L. — Waclawičová, M. — Křen, M. (2012): Balanced data repository of spontaneous spoken Czech. In: *Proceedings of LREC2012*, s. 3345–3349.

Waclawičová, M. — Křen, M. — Válková, L. (2009): Balanced corpus of informal spoken Czech: Compilation, design and findings. In: *Proceedings of INTERSPEECH 2009*. Brighton: ISCA, s. 1819–1822.

Lucie Benešová | Ústav Českého národního korpusu, FFUK | nám. J. Palacha 2, 116 38 Praha 1
lucie.benesova@ff.cuni.cz

Michal Křen | Ústav Českého národního korpusu, FFUK | nám. J. Palacha 2, 116 38 Praha 1
michal.kren@ff.cuni.cz

Martina Waclawičová | Ústav Českého národního korpusu, FFUK | nám. J. Palacha 2, 116 38 Praha 1
martina.waclawicova@ff.cuni.cz