TVORBA KORPUSŮ MLUVENÉHO JAZYKA

MILOSLAV VONDRÁČEK
Slezská univerzita v Opavě, Filozoficko-přírodovědecká fakulta v Opavě, Česká republika
miloslav.vondracek@fpf.slu.cz

Má někdejší univerzita se v minulých letech podílela na tvorbě korpusů mluvené komunikace. Spolu se studenty jsme pořídili zvukový záznam 220 soukromých dialogických situací a jejich přepis. Při té příležitosti jsme museli řešit řadu praktických problémů. Ty vedly k formulaci podstatných teoretických otázek. K základním patří relativita jednotek mluveného jazyka. Výsledkem je metodologie tvorby korpusu, od té doby neustále zdokonalovaná. Příspěvek přináší přehled základních otázek a snaží se poskytnout teoretické odpovědi i metodiku řešení.

DOI
https://doi.org/10.18690/um.ff.4.2024.11

ISBN
978-961-286-882-6

Ključne besede:
korpus mluveného jazyka, relativita jednotek řeči, pravidla přepisu, zvukový záznam, neoficiální komunikační situace

Univerz tetna založba
Univerze v Mariboru

DOI
https://doi.org/
10.18690/um.ff.4.2024.11

ISBN
978-961-286-882-6

Keywords:
spoken language corpora,
relativity of speech units,
transcription rules,
audio recording,
unofficial communication
situations

CREATION OF SPOKEN LANGUAGE CORPORA

MILOSLAV VONDRÁČEK
University of Ljubljana, Faculty of Arts, Ljubljana, Slovenia
mojca.smolej@ff.uni-lj.si

In recent years, my former university has been involved in the creation of corpora of spoken communication. The students made audio recordings of 220 private conversations. With my help, the students converted these dialogues into written text. On this occasion, we had to solve some practical problems. These difficulties led to the formulation of substantial theoretical questions. The relativity of the units of spoken language is one of the fundamental ones. The result is a corpus building methodology that has been continuously improved since then. The paper provides an overview of the fundamental questions and attempts to provide theoretical answers and a solution methodology.

University of Maribor Press

TVORBA GOVORJENEGA KORPUSA

MILOSLAV VONDRÁČEK
Šlezijska univerza v Opavi, Filozofska in naravoslovna fakulteta v Opavi, Republika Češka
miloslav.vondracek@fpf.slu.cz

V preteklih letih je moja nekdanja univerza sodelovala pri ustvarjanju korpusov govorjene komunikacije. Študenti so posneli 220 zasebnih pogovorov in jih z mojo pomočjo transkribirali. Ob tej priložnosti smo morali rešiti vrsto praktičnih težav. Te težave so privedle do oblikovanja pomembnih teoretičnih vprašanj. Eno temeljnih je relativnost enot govorjenega jezika. Rezultat tega je metodologija gradnje korpusa, ki se od takrat nenehno izboljšuje. Prispevek podaja pregled temeljnih vprašanj ter poskuša podati teoretične odgovore in metodologijo reševanja.

DOI
https://doi.org/10.18690/um.ff.4.2024.11

ISBN
978-961-286-882-6

Ključne besede:
korpusi govorjenega jezika, relativnost govornih enot, pravila transkripcije, zvočno snemanje, neuradne komunikacijske situacije

Univerzitetna založba Univerze v Mariboru

1 Úvod

Tvůrci korpusů mluvené komunikace řady ORAL, vznikajícího úsilím Ústavu Českého národního korpusu Filozofické fakulty Univerzity Karlovy v Praze, se zhruba v roce 2005 obrátili na další bohemistická univerzitní pracoviště České republiky s výzvou k účasti na jejich projektu.[1] Pracoviště už tou dobou mělo zkušenost s tvorbou Pražského mluveného korpusu (2001, dále PMK[2]), byly tu poznatky spjaté s Brněnským mluveným korpusem (2002, dále BMK[3]), z aktuálního sběru materiálu pro celočeský ORAL2006,[4] to vše na pozadí důkladné obeznámenosti s problematikou v mezinárodním měřítku. Spolupráce s mimopražskými středisky měla spočívat ve shromáždění zvukových záznamů neoficiálních komunikačních situací dostatečného rozsahu a jejich přepisu podle stanovených pravidel. Povaha spolupráce s organizačním centrem se mohla případ od případu lišit v závislosti na odborném (a nakonec i lidském) profilu prostředníka. Zpětně mohu konstatovat, že pro mne – a průkazně i pro angažované studenty[5] – se stala tato aktivita zdrojem širokého spektra poznatků ze všech lingvistických disciplín. To je důvod, proč neváhám zkušenost s odstupem a nadhledem uplynulých let znovu sdílet.[6]

2 Povaha komunikační situace

Korpusy ORAL2008 a ORAL2013,[7] k nimž se váže má akviziční zkušenost, jsou charakterizovány jako korpusy neformální mluvené češtiny. Sdružují „materiál představující prototypický spontánní mluvený jazyk, který se používá při bezprostřední interakci mluvčích v neformálních komunikačních situacích.

[1] Výsledky se promítají do korpusů ORAL2008 a ORAL 2013. K nim více na https://wiki.korpus.cz/doku.php/cnk:uvod

[2] Informace o PMK dostupné na https://wiki.korpus.cz/doku.php/cnk:pmk

[3] Informace o BMK dostupné na https://wiki.korpus.cz/doku.php/cnk:bmk. PMK i BMK ovšem zahrnují zčásti i moderované formální monologické promluvy.

[4] Informace o ORAL2006 dostupné na https://wiki.korpus.cz/doku.php/cnk:oral2006.

[5] Mé tehdejší působiště: Univerzita Hradec Králové, Pedagogická fakulta, roky 2005–2011. Zapojilo se postupně přes 200 studentů. Každý pořídil jedinečnou sondu do mluvené komunikace, tj. nahrávku o délce průměrně 30 minut (reálně 20–50 minut, celkem přes 100 hodin). Student sám svůj zvukový záznam přepsal (v první etapě do textového editoru Word, později do aplikace Transcriber). Přepis poté prošel několikanásobnou kontrolou věrnosti přepisu. V prvních kolech kontrolu zajišťoval příslušný akademický pracovník univerzity (na UHK já), finální kontrolu pracovník ÚČNK FF UK. U obou spolutvořených korpusů viz úsek Poděkování (https://wiki.korpus.cz/doku.php/cnk:oral2008, https://wiki.korpus.cz/doku.php/cnk:oral2013).

[6] O aktuálních zkušenostech z probíhajícího projektu jsem hovořil na konferenci Čestina v mluveném korpusu, Praha 2007, pro více informací viz Vondráček 2008.

[7] Oba ve srovnání s korpusem ORAL2006 kromě Čech zahrnují i oblast Moravy a Slezska (což bylo jedním z motivů k oslovení mimopražských univerzitních pracovišť).

Hlavními kritérii pro získávání nahrávek byly: fyzická přítomnost všech mluvčích na jednom místě, dialogičnost promluv, vzájemný blízký vztah mluvčích, nepřipravenost, spontánnost, neveřejná a neoficiální komunikační situace“ (ORAL2013, kráceno).

Starší korpusy popisují neformální komunikační situace dalšími znaky: „neformální promluvy tvoří dialogy dvou, případně i více mluvčích, kteří se dobře znají“ (BMK), resp. „kteří se znají“ (PMK). V případě korpusu ORAL2006 „[v]šechny nahrávky vznikaly výhradně v neformálních situacích, což znamená, že se mluvčí vzájemně znali a měli k sobě přátelský vztah.“ Při tomto zadání se projevila nejednoznačnost v interpretaci vykání (i mezi rodinnými příslušníky). Současně se objevil jiný nečekaný situační rys – účastník komunikace je sice v přátelském vztahu ke komunikačnímu partnerovi, současně však hovoří z titulu své profese (kamarádka kadeřnice při úpravě účesu, kamarád automechanik při sjednávání opravy vozu). Průnikem obou eventualit vzniklo další pomocné kritérium, v dostupných podkladech nezaznamenané: vykání samo o sobě není v rozporu s neformálností a neoficiálností komunikační situace; současně (bez ohledu na tykání) žádný z účastníků nevystupuje v dialogu z titulu své profese. Tato charakteristika se více než osvědčila při jemném rozlišení komunikačních situací v rámci stylistických analýz komunikační sféry běžné.

3 Delimitace syntaktických jednotek

Pro první přepisy záznamů (pořizovaných již na digitální záznamníky) jsme využívali textového editoru Word.[8] Ještě pro předchozí korpus ORAL2006 platila pravidla stanovení hranic syntaktických celků víceméně odpovídající zvyklostem psaných textů.[9] Mírný posun přináší korpus ORAL2008 ve způsobu zaznamenání obvyklých defektů mluveného projevu.[10]

[8] Složitou cestu uvědomování si a dokumentace rozdílů psané a mluvené češtiny ilustruje text J. Hoffmannové a M. Mikulecké (2011, 78–92).

[9] „Hranice vět se vyznačují jen interpunkcí, na začátku věty píšeme malé písmeno. Větné interpunkce se užívá tak, jak je to obvyklé v textech psaných, tj. nezachycuje se přerušování věty pauzami, naopak náležitou čárku ve větách a v souvětích píšeme, i když se větné předěly pauzou nerealizují. Neukončené věty označujeme třemi tečkami s dvojtečkou odsazenými mezerou, tedy …: Případ, kdy je mluvčí přerušen jiným mluvčím, ale ve výpovědi později pokračuje, se v dialogu značí třemi tečkami na konci přerušené výpovědi i na začátku její navazující části.“ (archiv autora, viz též dále) „Pořizování nahrávek, jejich přepisování a označování probíhalo v souladu s obecnými zásadami uplatňovanými při přípravě všech předchozích mluvených korpusů v rámci Českého národního korpusu, zejména korpusu ORAL2006.“ (https://wiki.korpus.cz/doku.php/cnk:oral2008)

[10] „Zvláštní rysy syntaktické stránky mluvených projevů (přeřeknutí, přerušování a změny větné perspektivy, přiřazování vět a větných úseků apod.) jsou zachyceny zjednodušeně, většinou pouze pomocí čárky. Pokud se slova opakují, jsou oddělena čárkou.“ (ORAL2008)

<3> dovopravdy funguje.
<2> já vim, nó, my sme jak já, já, né my, co budu říkat my, já sem za vola.
<3> né, to neni, neni, já ti říkám, čim to bylo, vy ste podle mě tady s tim šabo, šibovali a naklonilo se to, udělala se bublina a to je jedna z alternativ, která byla možná, a pře, pak to nechladilo, protože si to musí sednout.
<2> ale vono, já mam takovej pocit, že s tim předtim jakoby nikdo nehejbal a najednou to přestalo jít.
<3> druhá varianta je, že, že se, žes v podstatě se to dostalo do ňáký mezipolohy a že, eee, třeba při tom přepínači, což mohlo bejt, a že začal blbnout termostat. a znova začal fungovat. z ňákejch důvodů.
<2> hm, hm.
<0> to je nejspíš ...:
<2> to je, to sou, todle, Jarečku, viš, to mi vysvětli, víš, že já se tady ztrapňuju, rok tady máme ledničku vypnutou ... (úryvek ze sondy 05H006N)

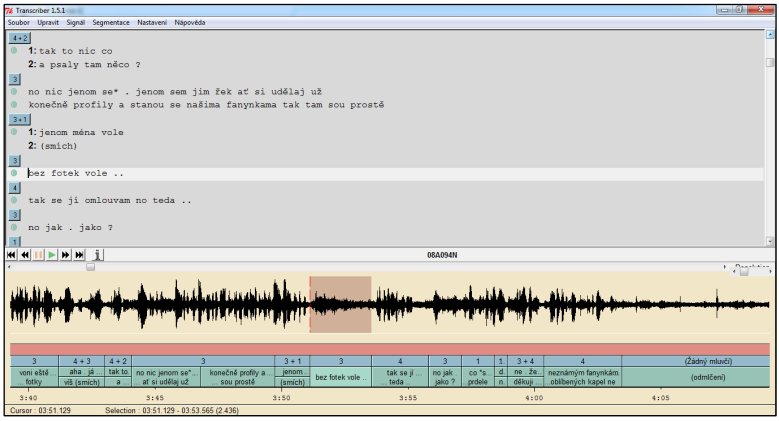
Počínaje korpusem ORAL2013 se přepis odehrává v transkripčním programu Transcriber.[11] To s sebou přináší – i vzhledem ke změně techniky převodu, opoře ve zrakové kontrole zvukového záznamu, odlišné segmentaci toku řeči[12] – změnu vnímání toku dialogu:[13] systematičtěji a komunikačně velmi přínosně se sledují překryvy komunikantů[14] (opouští se tedy snaha rozplést je do zdánlivě samostatných, nesimultánních replik), a zejména interpunkce převzatá z psaných textů je nahrazena interpunkcí pauzovou. Odlišují se tři kvality pauzy: pauza krátká, pauza delší a odmlčení; rozlišeny jsou podle individuálního tempa jednotlivých mluvčích.

[11] Program Transcriber neslouží k automatickému převodu zvukového záznamu na psaný text. Po uložení nahrávky ale přepisovateli umožňuje členit text do krátkých sekvencí podle kritérií formálních (max. délka, viz dále) nebo funkčních (replika jednoho mluvčího, pasáž překrývání více mluvčích ap.) Program sloužil jen pro přepis a kontrolu. Hotová, zkontrolovaná sonda převzatá pro zařazení do korpusu se dále zobrazuje v jiném nástroji (KonText, https://www.korpus.cz/kontext/query?corpname=syn2020, viz obr. 2) Všechny sondy vytvořené na Univerzitě Hradec Králové pod mou patronací mají na třetí pozici sedmimístné značky sondy písmeho H (př.: 09H015N = sonda vytvořená v roce 2009 v Hradci Králové jako 15. v pořadí, komunikační situace neoficiální).

[12] „[r]epliky jsou členěny na segmenty představující sémanticky, prozodicky i syntagmaticky ucelenou sekvenci v průměru o 5–10 slovech (maximálně však 15).“ (ORAL2013)

[13] K vertikální ose komunikátu viz kupř. podnětnou studii Nepravá hypotaxe v spontánních mluvených projevech (Bílková 2021)

[14] K tomu např. Komrsková–Poukarová 2018, 41–56; Komrsková–Poukarová–Havlík 2019, 102–116.



Obrázek 1: Přepis zvukového záznamu v programu Transcriber
Zdroj: https://wiki.korpus.cz/doku.php/cnk:oral2013. Printscreeny aplikace s přepisem studentských sond jsem nepořizoval

Tradiční interpunkcí se vyznačuje pouze nápadně stoupavá intonace věty tázací, popř. vykřičníkem nápadná zvolací intonace. Důvodem rezignace na standardní značení mezivětných předělů bylo zejm. stanovisko odborníků z oblasti fonetiky a fonologie,[15] že zvukové hranice takových vyšších textových jednotek lze nalézt jen stěží. Opakovaně se nám tak potvrzuje, že sémaziologická dekompozice textu probíhá primárně na základě kritérií významových.

Tabulka 1: Ukázka přepisu dialogu s pauzovou interpunkcí

Radomíra_7878	–	*to uplně cejtím že mně to chybí ty jo su uplně taková ..*
Adéla_5592	–	*no já taky . chcu jako něco . pořádně .*
Adéla_5592 + Radomíra_7878	– –	*když -* *mně hlavně .*
Radomíra_7878	–	*já hlavně cejtim jak sem závislá na Dimim . až teď s* dycky když jakoby sem vod něj .*
Adéla_5592	–	*@*
Radomíra_7878	–	*teď jak sem sem jela tak sem říkala . ty jo mně se sem nechce .*
Adéla_5592	–	*(smích)*
Radomíra_7878	–	*najednou musím všechno sama víš všechno*
Adéla_5592 + Radomíra_7878	– –	*no jasné no* *tak jakože ono ale*
Radomíra_7878	–	*mně to nepříde*

Vir: Zdroj: Korpus ORALv1, sonda 11A085N, https://www.korpus.cz/

[15] Fonetický ústav Filozofické fakulty Univerzity Karlovy Praha, tehdy pod vedením prof. PhDr. Zdenky Palkové, CSc. (https://fonetika.ff.cuni.cz/ustav/vyucujici/zdena-palkova/)

4 Fenomén lingvální, paralingální a nonlingvální

Nikoli nezajímavé je (a bylo nejen pro studenty filologických oborů) ověřování hranic živlu lingválního a nonlingválního s přechodovou oblastí v jevech povahy paralingvální. Pravidla platná pro korpus ORAL2006 stanovila dosti jednoduše, že parazitní zvukové projevy se zaznamenávají sekvencí tří písmen: obvykle *hmm* pro zvuky spíše souhláskové, *eee* pro zvuky povahy převážně vokalické. Smích se vyznačuje poznámkou v závorce (smích); v případě verbalizovaného, resp. i verbalizovatelného smíchu se zapisuje zvuková forma „co nejblíže slyšenému“.[16]

Už v průběhu sběru jazykového materiálu pro korpus ORAL2008 jsme narazili na nezbytnost odlišení komunikačně relevantních neartikulovaných nebo poloartikulovaných zvuků vyjadřujících souhlas, nesouhlas, pochybnosti, váhání apod. Zejména šlo o laryngálně-nazální zvuk (provázený případně odmítavým pohybem hlavou), graficky snad zaznamenatelný jako kombinace rázu a písmena pro konsonant *m*, tj. jako *´m-´m*. (Dlužno říci, že např. ani autoři Příruční mluvnice češtiny, dále PMČ, nečinili nijak podstatného rozdílu mezi výrazy *hm, ehm* a *mhm*, řadíce je souběžně k interjekcím (první dva) a partikulím (třetí), obecně přitakacím (PMČ 1995, 356 - § 593, a 365 - § 606). Garanti nově vznikajícího korpusu pochopili důležitost signalizace zásadně odlišné komunikační funkce prostředků hraniční povahy a stanovili další, strukturovanější pravidla transkripce: *hmm* pro přitakací responzní zvuk, *emem* pro nesouhlasný responzní zvuk, *mmm* pro souhláskové hezitační zvuky a *eee* pro zvuky hezitační samohláskové.[17]

Nejde o jev funkčně triviální. Při tvorbě hesla *ehm* pro Akademický slovník spisovné češtiny (dále ASSČ) odlišila Jana Špirudová[18] sedm jeho různých komunikačních platností (v dnešní zveřejněné verzi jsou dostupné čtyři, pročež uvádím kompletní znění hesla v neoklešněné podobě):

ehm

– V1 vyjádření nejistoty, váhání, nerozhodnosti, mírných rozpaků
Ehm, stala se taková věc…; Já jsem, ehm – zkrátka, nezlobte se, že jsem vás tak otravoval.; Ehm, jak ti to říct, zkrátka kluci nevěděli, že ta bonboniéra není naše.; Ehm, nevím, vážně mě to nijak extra neláká.; „Mohl bych… Ehm, mohl bych vás o něco požádat?“ vykoktal.

[16] Zásady přepisu pro korpus ORAL2006. Dostupné z: https://wiki.korpus.cz/doku.php/seznamy: pravidla_2006
[17] Inspirativní je zpracování výrazu *no* jako diskurzního markeru (Bílková-Zeman 2020, 191-198)
[18] Jako spoluautor ASSČ a osoba pověřená kontrolou tohoto hesla vidím informace o zakladateli (zpracovateli) hesla a autorech jeho vyšších verzí. Protože interjekce představují jednu z mých gramatických specializací, Jana Špirudová mě o revizi tohoto hesla výslovně požádala.

– *V2* opatrný, váhavý souhlas, kladná odpověď na otázku
Dáte si pivo? – Ehm, díky.; Ehm, tak to můžu říct docela přesně.

– *V3* zdůraznění otázky, pobídka k odpovědi
Ehm, nevadí ti, když půjdu s vámi?; Ehm a jak byste sestavil rozpočet vy?; Mohu vám, ehm, ehm, nějak pomoci?

– *V4* zvuk při odkašlání, kterým chce zprav. někdo na něco upozornit (např. že se chystá něco říct)
„Ehm," odkašlal si ředitel.; „Ehm, ehm, ehm," zakašlala paní Renata a kopla ho pod stolem prudce do holeně.; „Ehm, ehm," odkašlala jsem si významně, jak to mám ve zvyku před klíčovým projevem na jednání s komplikovaným partnerem.

– *V5* výraz zamyšlení nebo snahy posoudit návrh, event. rozvinout téma
Ehm, co kdybychom šli radši dovnitř, je tu celkem zima.; Ehm, ehm...ještě jedna věc.; Ehm – a co takhle použít polarizační filtr?

– *V6* výraz nelibosti, nesouhlasu nebo mírného odmítnutí
Ehm, to teda děkuju.; Ehm, to jste myslel vážně, pane? Ehm, to jako nemám říkat?
V7 povzdechnutí, vyjádření mírné námitky
Ehm, a je to tady zase.; Ehm, tak to už tady jednou bylo.

V úzu se vedle základní podoby citoslovce *ehm* objevují varianty opakující písmeno *m*, čímž se naznačuje delší trvání zvuku a zesiluje se tak komunikační účinek: *ehmm*, *ehmmm*. (ASSČ, heslo ehm, verze z 26. 5. 2022, před umrtvením významů 5-7)

Tabulka 2: Frekvence grafických forem responzních a hezitačních zvuků

	Grafická forma	Absolutní frekvence[19]	Frekvence na milion slov (i. p. m.)
ORALv1	*hmm*	50 095	7 874,46
	emm	636	99,97
	ééé	5	0,79
	em em	1	0,16
	chm	1	0,16
PMK	*hm*	760	927,66
	ehm	143	174,55
	eee	60	73,24
	hmm	35	42,72
	chm	27	32,96
	mmm	25	30,52
	ééé	4	4,88
	emem	3	3,66
	em em	1	1,22
	ehmm	1	1,22
ORTOFON	*hmm*	29 200	11 403,62

[19] Absolutní frekvence je ovlivněna velikostí konkrétního korpusu. Parametr i. p. m. udává přepočtenou frekvenci na ideální jeden milion slov daného korpusu.

	Grafická forma	Absolutní frekvence[19]	Frekvence na milion slov (i. p. m.)
	emm	579	226,12
	@ / @@	23 606	9 218,97
	ehm	6	2,34
	eee	3	1,17
	mmm	1	0,39

Zdroj: vlastní

S ohledem na budoucí lemmatizaci byla cíleně sjednocována přirozeně oscilující forma zápisu (respektující nakonec pokyn zapisovat co nejvěrněji slyšené). Logicky se tak potlačuje možnost vyjádřit počtem grafémů či jejich skladbou délku či obecnou povahu zvuku. Do transkriptů se však přece jen individuální reflexe v jisté míře prosadila.

Současná praxe dvouúrovňového přepisu, uplatňovaná pro korpus ORTOFON, nahlíží tuto oblast komunikace zase jiným prizmatem. Pravidla[20] ukládají přepisovat na rovině ortografické responzní zvuky přitakávací *hmm,* responzní zvuky odmítací *emm*, citoslovce bez lexikalizované podoby jako *&* a hezitační zvuky znakem *@* pro kratší hezitace a *@@* pro delší. Není zřejmé, jestli ojedinělé odlišné zaznamenané formy vypovídají o snaze vědomě zachytit specifický jev přesahující pravidla, nebo jsou jen projevem přehlédnutí. (Pro srovnání: Korpusy psaného jazyka dokládají, že psané texty, zejm. beletristické a publicistické, pracují i s dalšími formami, v mluvených korpusech registrované nanejvýš okrajově; v SYNv11 kupř. *chm* 561x / 0,09 i.p.m., *chmm* 131x / 0,02 i.p.m., *chmmm* 4x, *chmmmmm* 1x. To může odpovídat komunikačně funkčnímu, převážně nazálnímu nelaryngálnímu výdechovému zvuku, povzdechu se sevřenými rty.)

> *Víte ... **chm** ... zrovna než jste přišel ... prohlížel jsem podložní sklíčka ... a přišel jsem najedno, které by vás mohlo zajímat.* (korpus SYNv11)

> *. --- brácha a tatinek si dycky přál syna a syna teda měl . i když se narodil později než já . ale dycky já sem byla taková .. **mmm** .. no nechci říct úplně v koutku . ale Jirka dycky moh víc ---- (korpus ORALv1)*

[20] Transkripce v korpusu ORTOFON. Dostupné z: https://wiki.korpus.cz/doku.php/cnk:ortofon: pravidla

Tabulka 3: Ukázka přepisu dialogu s pauzovou interpunkcí a responzními zvuky

Ivana Š.	–	*hmm*
Rozálie B.	–	*víš ? .. hele fakt do mě nic nebylo hele Ilonko*
Ivana Š.	–	*hmm ..*
Ivana Š.	–	*hmm .. hmm .. hmm*
Rozálie B.	–	*do mě nic nebylo jo*
Ivana Š.	–	*hmm*
Rozálie B.	–	*no a naštěstí teda když přijeli ..*
Rozálie B.	–	*tak mi řekli že to byl teror že je . v devět večer přivezli na pokoje*
Rozálie B.	–	*a v sedm ráno vstávali protože se zase jede ne ..*
Ivana Š.	–	*jasně byli urvaný*

Zdroj: Korpus ORTOFONv2, sonda 18X096N, https://www.korpus.cz/

Tabulka 4: Ukázka přepisu s nelexikalizovanými interjekcemi (&) a hezitačními zvuky (@)

Anna V.	–	*tady v Brně má @ Kokino se to jmenuje je to značka . to dělá nějaká slečna a ..*
Anna V.	–	*má tam @ čokodortík s mořskou solí měla #s to někdy ?*
Miriam Š.	–	*ehm*
Anna V.	–	*jestli máš ráda čokoládku a karamel .. z tohohle by ses úplně potento protože -*
Miriam Š.	–	*musíš dát potom nějaký odkaz*
Anna V.	–	&
Anna V.	–	& *. to je víš kde to mají ? . to mají na Gor* @ Gorkého . víš kde*

Zdroj: Korpus ORTOFONv2, sonda 14A004N, https://www.korpus.cz/

5 Delimitace lexikálních jednotek a jejich hláskové struktury

S relativitou textových jednotek na úrovni věty a souvětí, resp. výpovědi, jakož i s otázkou hranic jevů lingválních a nonlingválních souvisí relativita centrální jednotky lexikálního systému (a svým způsobem centrální jednotka jazyka vůbec), totiž slova. Znovu platí, že vyčlenění jednotky v proudu řeči je značně relativní, formální delimitace vůči okolním jednotkám pauzou je hypotetická. Přirozený jazyk ze své podstaty inklinuje k jednotě formy a funkce, což posiluje tendenci k formálnímu splývání původně analytických jednotek. Podobně vede změna funkcí k proměnám formy. Všechny náznaky těchto procesů v mluvené komunikaci má proto smysl registrovat.[21]

[21] K tomu viz např. Blatná 2006, 7–19, Vondráček 2013, 79–89, Volín 2015, 44–47, Vondráček 2018, 35–110.

Tak by bylo možno sledovat ukázkovou proměnu desubstantivních forem kontaktového prostředku citoslovečného a částicového rázu *ty vole* (chápaného nejen ve školském prostředí funkčně neadekvátně jako vulgární, resp. zhrubělé[22] oslovování) ve zvukové a grafické realizaci: jeho hláskovou variabilitu či erozi včetně intonačního, resp. obecně prozodického pozadí v proměnách doby a v závislosti na funkci (→ part. *tyvoe* → *tyoe* → *tye;* → interj. *tývoe* → *týe* vedle snad tabuové náhražky *týjo* → *týo*). Z dobrých důvodů ovšem tematizuji samostatně dva jiné jevy podobné povahy.

Důvody pro upuštění od detailní analýzy jevu již naznačeného (deskripce slovnědruhové transpozice kontaktové fráze spočívající v pronominu + vokativu substantiva *ty vole* v kontaktovou partikuli *tye* či podivovou interjekci *týe*) jsou naznačené povahy axiologické (byť příznak vulgárnosti či zhrubělosti v jazykovém materiálu nebyl v žádném případě důvodem k vyřazení sondy). Volím však dva jiné, analogické, ovšem různě složité případy funkční a slovnědruhové proměny provázené destrukcí formy: proměnu substantivního vokativu *člověče* v jednotku pronominálního typu *čék*, v kontaktovou partikuli *čéče* a v podivovou interjekci *čéče*. Dílem jiné (tabuové, ovšem v tomto případě nábožensky tabuové) jsou pohnutky k obměnám forem (a k jejich následným deformacím) u interjekcí s funkcí povzdechu či zaklení typu *proboha, prokrindapána*. Podobných ilustračních typově odlišných dokladů je možno zvolit desítky; uvádím právě jen některé z těch, jež si obvykle i v přepisovacích pravidlech vyžádaly samostatné pojednání. K obému viz např. Laubeová 2020.

5.1 Typ *člověk / čék, člověče / čoveče / čéče*

Starší české slovníky zaznamenávají výraz *člověče* jako vokativ substantiva, funkčně jej označují jako expresivní oslovení, zařazují jej k hovorové nebo obecné češtině. ASSČ uvádí původní vokativní formu jako samostatné heslo, slovnědruhově interpretované jako interjekci (kolokviální vyšší), k ní dále i variantní formy *čoveče, čéče* jako kolokviální. Právě o zaznamenání hláskové eroze slova při slovnědruhové transpozici nám jde (a zmínit je třeba i funkci částicovou, v níž výraz nevykazuje znaky větného ekvivalentu). ASSČ také pod substantivním heslem *člověk* uvádí uplatnění ve funkci zájmena neurčitého, záporného nebo osobního s významem 'blíže neurčená osoba, někdo, kdokoliv, každý, mluvčí (já), adresát (ty, vy) apod., ve

[22] SSJČ i SSČ svorně: zhrub., i nadávka, hlupák; kontaktovou frázi *ty vole* však tyto zdroje ani IJP neuvádějí.

spojení se záporným slovesem žádná osoba, nikdo'. I tato slovnědruhová transpozice je v hovoru provázena destrukcí formy (*čék*).

Obrázek 2: Ukázka konkordancí s výrazem *čéče* ve webovém rozhraní KonText

Následující tabulka ukazuje, které z forem zachytil který z korpusů, popř. jak jsou morfologicky značkovány.

Tabulka 5: Korpusově evidované formy výrazů *člověk / čék, člověče / čoveče / čéče*

	Grafická forma	Absolutní frekvence	Frekvence na milion slov (i. p. m.)
PMK	*čoveče / čověče*	49 / 3	59,81
PMK	*člověče*	15	18,31
PMK	*čoeče*	1	1,22
PMK	*čék*	1	1,22
ORAL2008	*čoveče / čověče*	130 / 14	96,33
ORAL2008	*člověče*	16	11,86
ORAL2008	*čéče*	2	1,48
ORAL2008	*čék*	2	1,48
ORAL2013	*čoveče / čověče*	547 / 10	166,49
ORAL2013	*člověče*	16	4,87
ORAL2013	*čéče*	17	5,17
ORAL2013	*čék*	0	0
ORTOFONv2	*čoveče / čověče* (subst.)	12 / 0	4,69
ORTOFONv2	*člověče* (subst.)	233	90,99
ORTOFONv2	*čéče* (subst.)	0	0
ORTOFONv2	*čék* (subst.)	0	0

Zdroj: vlastní

... --- sou takový ještě vomámený ty ryby , že jo ? ... povidal, to sem ale m ...: --- voni sou přiblbý,* ***čoveče****, voni vůbec nejdou, voni zůstanou u břehu v tý trávě a vůbec nikam nejdou . no* (korpus ORAL2008)

už sem nemohla . no a já říkám . zas to máš blbě . ***čéče*** *co t* jako něco takovýho no a . tam je jedna taková šíleně . mmm . šílená poděska . a vona hotová uplně jako .* ***čéče*** *jo a toto . a teď se tam řehtali . hroznýý* (korpus ORAL2013)

... no prostě, stálo jim to za to, no . jesi ... : no pokavaď, že jo, pokavaď je to fyzicky a duševně vodrovná, tak ***čék*** *nemůže říct: stálo to za to, no . holt, já nevim, no .* (korpus BMK)

... v noci v noci dyž se ***čék*** *probudí tak to bolí, a nevim jak si lehnout, fakt je že pak si nebudu smět dát nohy křížem že jo třebas, vsedě určitě ne, jo jak s** ***ček*** *normálně sedí a přehodí si nohu tak* (korpus ORALv1)

Stejně jako koreluje změna funkce slovní formy s omezením či ztrátou paradigmatu a s erozí fonetické struktury, dochází ke změnám u víceslovných lexikálních jednotek (historicky nejspíš takto haplologicky vznikla slova *vašnosti* ← *Vaše Jasnosti* s dodatečným Nsg *vašnosta, slečna* ← *šlechtična*, týmž směrem se ubírala forma *pančelka* ← *paní učitelka*, v mírné podobě nakonec i dosud nekodifikované *nashledanou* ← *na shledanou, *shledaná*). Také toto splývání forem a redukční procesy má smysl sledovat. I pro druhé uvedu ilustrační příklad.

5.2 Typ *proboha / probůh, prokristapána, propánajána, propánaboha*

Všechny velké slovníky češtiny 20. století registrují lexikální formu *proboha* (některé i *probůh*) jako citoslovce;[23] neuvádějí varianty psané analyticky (tudíž nemusí řešit pravopis velkých písmen vázaný na tuto funkci). Lístková excerpta[24] tvořící podklad Příručního slovníku jazyka českého (dále PSJČ) registrují ve třinácti dokladech i lexikální formy *prokristapána* a jeho tabuovou obměnu *prokrindapána*, vedle ní s nižším počtem dokladů i *propánajána, propánaboha,* vše hodnoceno jako interjekce (slovníková hesla ovšem nevznikla). Pozdější slovníky už heslo uvádějí (SSČ[25] *prokrista(pána)* i *pro krista pána*, SSJČ[26] *prokrista, prokristapána,* psáno též *pro krista pána*, obdobně *propána, propánajána, propánaboha, propánakrále* (psáno též *pro pána boha*), rovněž jako interjekci. Psaní malého písmena lze přiřkout spíše režimním pohnutkám než snaze odlišit rouhavé braní jména Syna Božího nadarmo od vroucího obvolávání se na něj; proti tomuto směru uvažování naopak hovoří připouštěné analytické psaní. Nezdá se nepodstatné, jak je týž výraz zaznamenán v přepisech pro mluvené korpusy sto let od vzniku PSJČ.

[23] Srov. též např. Kleňhová 2010; 2012, 238–254.
[24] Elektronicky jsou dostupná z webu Ústavu pro jazyk český AV ČR, v. v. i.: https://psjc.ujc.cas.cz?
[25] *Slovník spisovné češtiny pro školu a veřejnost*, 1994. Praha: Academia.
[26] *Slovník spisovného jazyka českého*, 1989. Praha: Academia.

Tabulka 5: Korpusově evidované formy výrazů *prokristapána / propánajána / propánaboha*

	Grafická forma	Absolutní frekvence	Frekvence na milion slov (i. p. m.)
PMK	*proboha / probůh*	7 / 0	8,54 / 0
PMK	*pro boha*	0	0
PMK	*prokristapána / propánajána / propánaboha*	3 / 0 / 0	3,66 / 0 / 0
PMK	*pro krista pána / pro pána jána / pro pána boha*	0 / 0 / 0	0 / 0 / 0
BMK	*proboha / probůh*	9 / 1	15,1 / 1,68
BMK	*pro boha*	0	0
BMK	*prokristapána / propánajána / propánaboha*	0 / 0 / 0	0 / 0 / 0
BMK	*pro krista pána / pro pána jána / pro pána boha*	0 / 0 / 0	0 / 0 / 0
ORAL2008	*proboha / probůh*	6 / 0	4,45 / 0
ORAL2008	*pro boha*	0	0
ORAL2008	*prokristapána / propánajána / propánaboha*	4 / 0 / 0	2,96 / 0 / 0
ORAL2008	*pro krista pána / pro pána jána / pro pána boha*	0 / 0 / 0	0 / 0 / 0
ORAL2013	*proboha / probůh*	53 / 0	16,13 / 0
ORAL2013	*pro boha*	7*	5,04
ORAL2013	*prokristapána / propánajána / propánaboha*	0 / 1 / 0	0 / 0,3 / 0
ORAL2013	*pro krista pána / pro pána jána / pro pána boha*	0 / 0 / 0	0 / 0 / 0
ORTOFONv2	*proboha / probůh*	23 / 2	8,98 / 0,78
ORTOFONv2	*pro boha*	1*	0,39
ORTOFONv2	*pro krista pána / pro pána jána / pro pána boha*	2 / 0 / 0	0,78 / 0 / 0
ORTOFONv2	*prokristapána / propánajána / propánaboha*	1 / 0 / 0	0,39 / 0 / 0

* (také) v adekvátní funkci předložkové substantivní vazby
Zdroj: vlastní

.. každopádně si mysím že by ženy se svými vědomostmi . @ . at už životními nebo ze škol . neměly zůstat takzvaně na ocet ale ***probůh*** *už ne žádnou další Anežku Hodinovou Vzpurnou výstřelky . @ . byly vždycky . @ . no* (BMK)

mě zamkly v pokoji .. co si vůbec myslely .. že jsem .. jo a Pavel říkal no když #s jim utíkala .. tak . pr .* ***prokristapána*** *copak měli dělat .. sestřička tě honila po baráku a nemohla tě najít .. tak ji zamkli ..* (ORTOFONv2)

. a jo , vlastně , já sem jí slíbila tu kaši, ***prokristapána*** *. čekáš, až ti to vystydne?*
ne, už du . już odcházím . bože , bože . no , podívej . eee . (ORAL2008)

lusev + pujova *a samý placení že jo*

jo právě
lusev + pujova *to stojí peněz potom* ***ježkovy voči*** *tak*
to je jistý to je pe peněz viď*
lusev + pujova *j* já jim do toho mluvit nebudu* ***propánajána*** *--- ňáký*
jo to jako jistý (ORAL2013)

voduna *no to víš no tak --- nevíš jaká budeš ty* (smích)
rehega *doufám že takovádle ne*
voduna (se smíchem) *no to nevíš tak jako*
rehega ***pro boha svatýho***
voduna (smích) *no a tak jako* (ORAL2013)

Světlana R. .. *tak tam byla nějaká mladá holka myslím si že nějaká středoškolačka že se zaučovala ne? .. učila se*
Růžena J. (pousmání) *jo .. ne tak to já bych asi nechtěla aby ta se zaučovala na mně* ***probůh***
Světlana R. (smích) *no .. a právě .. přede mnou teda byla nějaká holka .. nejdřív brali z prstu to už ti brali krev z prstu .. taky? .. nebo ne?* (ORTOFONv2)

V zájmu automatizovatelné lemmatizace a snazšího vyhledávání vychází dvouúrovňový korpus ORTOFONv2 z ortgografického přepisu; v zásadě proto není divu, že emocionální výrazy zaznamenává především analyticky. Jeho (popředložkové) komponenty jsou však tagovány jako substantiva. Tím se zcela stírá rozdíl mezi interjekcí a substantivním užitím, doloženým v témže korpusu kupř. sekvencí:

...to jsou .. to je ***pohled*** *prostě* ***pro boha*** *jako tak jo jakože jo takové jako přirovnání jakože ..* (ORTOFONv2)

Tuto pasáž, věnovanou změnám forem a funkcí prostředků mluvené komunikace, jsem uvedl příkladem spojení *ty vole* / *tyoe* / *tyve* / *tye*, užívaného mj. jako kontaktová partikule. Při sběru materiálu pro korpusy, na nichž jsme spolupracovali, jsem považoval za potřebné všechny podobné náznaky funkčních změn zaznamenávat. Poslední zmíněný korpus ORTOFONv2 eviduje 2630 dokladů formy *vole* a jedinou formu *voe*. Všechny ortograficky ztvárněné výrazy jsou tagovány jako substantiva. Tímto postupem je zcela zastřena funkční diference partikule, interjekce a jména (jako takového významně frekvenčně omezeného):

Zbyněk S. *hochu ale co mě překvapil třeba Jenda NP .. tam byl a .. ty jo ale já tomu Babišovi fandím ..*
Norbert R. (smích) *kokote jeden* ***vole*** *že to tak řeknu ..*

6 Závěr

Pokusil jsem se shrnout několik zásadních, dosti různorodých momentů, jež jsme seznali za hodné pozornosti při tvorbě mluvených korpusů s pomocí studentů filologických oborů. Ti, nakonec stejně jako my sami jsme se při obstarávání zvukových záznamů neformálních komunikčních situací a jejich přepisu museli zabývat značně obecnými otázkami pragmalingvistickými a sociolingvistickými (povaha komunikační situace), obecně lingvistickými (sémioticky heterogenní povaha komunikátu, počítaje v to prolínání prostředků lingválních, paralingválních a nonlingválních), lexikologickými a morfosyntaktickými (delimitace jednotek v proudu řeči a zachycení jejich relevantní formy).

Porovnáním více korpusů mluveného jazyka se snažím ukázat dynamiku sledování určitých jevů a zisky, popř. ztráty plynoucí ze změny intence. Dílčí omluvou budiž, že omezení prostoru pro individuální snahu transkriptora zaznamenat co nejvěrněji obraz zvukové formy je veden objektivizačním úsilím. Právě to bylo i pohnutkou k ústupu od interpunkce psaných textů a přechodu k interpunkci pauzové. Dalším zásadním motivem bude (resp. dlouhodobě je) strojové učení a automatické zpracování mluvené formy řeči.

Literatura

Jana BÍLKOVÁ, 2021: Nepravá hypotaxe v spontánních mluvených projevech. *Štýl – komunikácia – kultúra.* Ur. Zuzana Popovičová-Sedláčková. Bratislava: Univerzita Komenského. 417–427.

Jana BÍLKOVÁ, Jiří ZEMAN, 2021: Diskurzní marker no a v mluvené češtině. Lingvistika – korpus – empirie. Praha: Ústav pro jazyk český AV ČR, v. v. i., 191–198.

Renata BLATNÁ, 2006: *Víceslovné předložky v současné češtině.* Praha: NLN, Nakladatelství Lidové noviny: Ústav Českého národního korpusu, 9–17.

Miroslav GREPL, 1995: Příruční mluvnice češtiny. Praha: NLN, Nakladatelství Lidové noviny, 1995.

Jana HOFFMANNOVÁ, Marie MIKULOVÁ, 2011: Korpusy mluvené češtiny a možnosti jejich využití pro poznání rozdílných "světů" mluvenosti a psanosti. *Korpusová lingvistika Praha 2011 - 2 Výzkum a výstavba korpusů.* Ur. František Čermák. Praha: NLN, Nakladatelství Lidové noviny. 78–92.

Eliška KLEŇHOVÁ, 2010: Interjekce v českém jazykovém systému. Kvalifikační práce magisterská, Filozofická fakulta Univerzity Karlovy.

Eliška KLEŇHOVÁ, 2012: Postavení a užívání interjekcí v současné češtině. *Naše řeč,* roč. 95, č. 5, 238–254.

Zuzana KOMRSKOVÁ, Petra POUKAROVÁ, 2018: Kdo, kdy a proč skáče komu do řeči aneb překryvy ve spontánní mluvené češtině. *Korpus – gramatika – axiologie*, č. 17, 41–56.

Zuzana KOMRSKOVÁ, Petra POUKAROVÁ, Martin HAVLÍK, 2019: Překryvy replik. *Syntax mluvené češtiny.* Ur. Jana Hoffmannová, Jiři. Homoláč, Kamila Mrázková. Praha: Academia. 102–116.

Zuzana LAUBEOVÁ, 2020: *Mluvenost v dialogické elektronické komunikaci.* Kvalifikační práce dizertační,

Filozofická fakulta Univerzity Karlovy.
Jan VOLÍN, 2015: Vztah mluvené a psané formy jazyka. *Mluvnice současné češtiny. 1, Jak se píše a jak se mluví*. Ed. Václav. Cvrček. Praha: Univerzita Karlova, nakladatelství Karolinum. 44–47.
Miloslav VONDRÁČEK, 2008: Diskrétnost řečových jednotek. *Čeština v mluveném korpusu.* Ur. Marie Kopřivová, Martina Waclawičová. Praha: Nakladatelství Lidové noviny / Ústav českého národního korpusu. 255–261.
Miloslav VONDRÁČEK, 2013: Vlastnosti slova a slovní druhy. *Akademická gramatika spisovné češtiny*. Ur. František. Štícha. Praha: Academia, 2013. 79–89.
Miloslav VONDRÁČEK, 2018: Druhy slov. *Velká akademická gramatika spisovné češtiny*. Ur. František Štícha. Praha: Academia, 2018. 35–110.
Slovník spisovného jazyka českého, 1989. 2. vyd. Praha: Academia.
Slovník spisovné češtiny pro školu a veřejnost, 1994. 2. vyd., opr. a dopl. Praha: Academia.

Internetové zdroje:

ASSČ: Akademický slovník současné češtiny. Dostupný z: https://slovnikcestiny.cz/uvod.php. Poslední přístup 15. 6. 2023.
BMK: Brněnský mluvený korpus. Dostupný z: https://wiki.korpus.cz/doku.php/cnk:bmk. Poslední přístup 15. 6. 2023.
ORAL2006: Dostupný z: https://wiki.korpus.cz/doku.php/cnk:oral2006. Poslední přístup 15. 6. 2023.
ORAL2008: Dostupný z: https://wiki.korpus.cz/doku.php/cnk:oral2008. Poslední přístup 15. 6. 2023.
ORAL2013: Dostupný z: https://wiki.korpus.cz/doku.php/cnk:oral2013. Poslední přístup 15. 6. 2023.
ORTOFONv2: Dostupný z: https://wiki.korpus.cz/doku.php/cnk:ortofon. Poslední přístup 15. 6. 2023.
PMK: Pražský mluvený korpus. Dostupný z: https://wiki.korpus.cz/doku.php/cnk:pmk. Poslední přístup 15. 6. 2023.
PSJČ: *Příruční slovník jazyka českého. Kartotéka lexikálního archivu (1911–1991).* Dostupný z: https://psjc.ujc.cas.cz/search.php? Poslední přístup 15. 6. 2023.