Univerzita Karlova v Praze
Filozofická fakulta

Diplomová práce

2014 Bc. Karolína Vyskočilová

Univerzita Karlova v Praze
Filozofická fakulta
Ústav českého jazyka a teorie komunikace

Diplomová práce

Karolína Vyskočilová

Tvorba specializovaného korpusu banátské češtiny a jazyková analýza vybraných jevů

Building a Specialised Corpus of Banatian Czech with a Linguistic Analysis of Selected Features

Praha 2014

Vedoucí práce: Mgr. Eva Lehečková, Ph.D.

V první řadě děkuji Mgr. Evě Lehečkové, Ph.D., za vedení práce a vstřícnost, se kterou práci přijala.

Ráda bych také poděkovala kolegům z ÚČNK, zejména Mgr. Lucii Benešové, PhDr. Marii Kopřivové, Ph.D., Mgr. Michalu Křenovi, Ph.D. a Mgr. Pavlu Vondřičkovi, Ph.D, kteří v průběhu let stáli v pozadí tvorby korpusu BANÁT a pomohli k jeho zdárnému dokončení svými radami i zajištění technické části.

Dále bych chtěla poděkovat svým přátelům a kolegům za podnětné rady a ochotu konzultovat jakékoliv problémy, zejména Zuzaně Komrskové, Zuzaně Mickové, Jiřímu Miličkovi, Anně Morávkové a Evě Volenové.

Děkuji své rodině za plnou podporu v průběhu mého celého studia.

Nakonec děkuji celé vesnici Bígr.

Prohlašuji, že jsem diplomovou práci vypracovala samostatně, že jsem řádně citovala všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze dne 15. srpna 2014

ABSTRAKT

Předmětem této práce je tvorba specializovaného korpusu mluvené banátské češtiny.

V teoretické části je popsán historický kontext a jazyková situace na území Banátu s důrazem na stav od 90. let 20. století do současnosti. Následuje podrobné nastínění problematiky specializovaných jazykových korpusů se zaměřením na korpusy mluvené češtiny.

Podrobně je zdokumentována samotná tvorba specializovaného korpusu BANÁT (metoda sběru, zpracování a přepisu nahrávek; vytvoření korpusu a jeho zveřejnění).

Následný empirický výzkum a analýza příklonek, přivlastňovacích zájmen, záporných zájmen a zájmenných příslovcí a příklonek v bígerské mluvené češtině potvrdily statisticky signifikantní odlišnost od stavu současné češtiny na našem území (u všech sledovaných jevů kromě příklonek).

Klíčová slova

Banát, banátská čeština, Bígr, čeština v zahraničí, korpus BANÁT, mluvený jazyk, specializované korpusy, tvorba korpusu, zájmena

ABSTRACT

The purpose of this thesis was to build a specialised corpus of Banatian Czech.

The theoretical section describes the historical and language situation in the Banat area focusing on its development since the 1990s to the present day. This is followed by a presentation and description of specialised corpora focusing on a corpus of spoken Czech.

Building a specialised corpus named BANAT is documented in detail (method of collection, processing and transcription of records, building of corpora and publishing).

The analysis of possessive pronouns, negative pronouns and adverbs and enclitics confirms a statistically significant difference between Banatian Czech and the present state of Czech language (in all observed phenomena except enclitics).

Keywords

Banat, Banatian Czech, Bigr, corpus BANAT, corpus building, Czech language abroad, pronouns, specialised corpora, spoken language

OBSAH

SEZNAM GRAFŮ, MAP A TABULEK

SEZNAM ZKRATEK

AKCES	Akviziční korpusy českého jazyka (UČJTK, FF UK)
ČJA	Český jazykový atlas
ČNK	Český národní korpus (projekt ÚČNK)
FF UK	Filozofická fakulta Univerzity Karlovy
frek.	frekvence
ipm	*instances per million* - počet výskytů jevu na milion (vztaženo k celému korpusu)
konf.	konfidenční interval (interval spolehlivosti)
SSJČ	Slovník spisovného jazyka českého
ÚČJTK	Ústav českého jazyka a teorie komunikace FF UK
ÚČNK	Ústav Českého národního korpusu FF UK

1. ÚVOD

Předmětem této práce je zejména tvorba specializovaného korpusu banátské češtiny (BANÁT) a následná analýza vybraných jevů. Banátskou češtinou označuji jazyk, kterým mluví Češi žijící od 1. poloviny 19. století na území rumunského Banátu.[1] Do dnešní doby se v Banátu zachovalo šest vesnic s kompaktním českým osídlením, které si uchovaly tradice a jazyk po dvě staletí. Ve mnou vytvořeném korpusu prozatím banátskou češtinu reprezentují nahrávky z Bígru, jedné z českých vesnic. Situace v zahraničních jazykových ostrovech se mění poměrně rychle (zejména v důsledku reemigrace či odchodu za prací do rumunských vesnic a následné asimilace s majoritou), proto je třeba jejímu zkoumání věnovat pozornost, dokud je to možné.

Česká menšina je nejmenší uznávanou minoritou v Rumunsku. Početně se jedná o pouhou jednu setinu procenta obyvatel, oproti ostatním menšinám je však velmi kompaktní. Důvodem uzavřenosti české komunity není jen geografická izolovanost a v minulosti také jazyková bariéra, ale zejména náboženské a kulturní rozdíly. Rumunští Češi jsou katolíci (v případě části Svaté Heleny evangelíci), zatímco většinu rumunského obyvatelstva tvoří pravoslavní křesťané. Dlouhou dobu soudržnost komunity podporovalo i to, že krajané uzavírali sňatky pouze mezi sebou (Costachie et al. 2011: 8–10). V dnešní době je díky dlouhodobému působení diglosie česká menšina bilingvní, ale čeština si stále uchovává svoji prestiž.

Teoretická část práce stručně shrnuje historii českého osídlení v Banátu a její součástí je i přehled zkoumání jazykové situace na území rumunského Banátu s důrazem na vývoj od 90. let 20. století do současnosti. Tato část se dále věnuje problematice specializovaných jazykových korpusů, a to zvláště českých mluvených korpusů.

Těžištěm celého projektu je tvorba specializovaného korpusu banátské češtiny BANÁT, kterému je věnována další část této studie. Jednotlivé fáze jeho vzniku jsou v ní podrobně popsány a dokumentovány: metoda sběru materiálu, zpracování nových i starých nahrávek

[1] V této studii bude pojem „Banát“ či adjektivum „banátský“ vyhrazen pouze pro území rumunského Banátu, v užším smyslu pak pouze pro šest vesnic s kompaktním českým obyvatelstvem (Bígr, Eibentál, Gerník, Rovensko, Svatá Helena a Šumice).

pro tvorbu korpusu – od přepisu přes anotaci a značkování až po samotnou přípravu korpusu k publikaci.

V poslední části je nasbíraný materiál podroben lingvistické analýze. Zkoumána byla pozice příklonek, přivlastňovací zájmena, záporná zájmena a zájmenná příslovce, tj. jevy, které se zdály být charakteristické pro banátskou češtinu v předchozích výzkumech. Podklad pro srovnání bígerské češtiny se stavem na našem území tvoří korpus BANÁT vytvořený v rámci této práce a korpus mluvené češtiny ORAL2013 vybudovaný Ústavem Českého národního korpusu (ÚČNK).

Je vhodné podotknout, že tato práce v mnohém navazuje na můj předchozí výzkum (Vyskočilová 2012b, 2013, v tisku), který dále rozvíjí a doplňuje. Hlavní změnou je, že korpus BANÁT byl rozšířen téměř na dvojnásobek a opraven. V brzké době bude zpřístupněn i se zvukovou stopou širší veřejnosti (plánováno na podzim 2014). Druhá změna se váže k dříve nastíněné možnosti zkoumání odlišností banátské syntaxe, vybrané signifikantní jevy jsou zde rozpracovány do větší hloubky, než tomu bylo v bakalářské práci.

1.1. Stručně k historii banátských Čechů

V této podkapitole je v krátkosti shrnuta historie českého osídlení v rumunském Banátu s cílem ozřejmit problematiku českého jazykového ostrova v Banátu. Pro detailnější popis historie a kulturního vývoje lze doporučit následující literaturu sloužící zároveň jako zdroj této podkapitoly: Karas (1937: 12–48), Secká (1995), Salzmann (1984: 65–71), Schlögl (1925: 5–19), Urban (1930: 5–16), případně novou rumunskou souhrnnou publikaci (Gesce 2013), která vyšla i v českém překladu.

Historické území Banátu je dnes rozděleno mezi Maďarsko, Srbsko a Rumunsko. Dříve ale patřilo celé Habsburské monarchii a po roce 1739 bylo součástí Vojenské hranice. Nížinaté části Banátu byly osídleny krátce po získání území od Turků, ale nejjižnější, hustě zalesněnou, hornatou část východního Banátu začali osidlovat až čeští kolonisté v roce 1823. Důvodů pro přesídlení bylo několik: hlavní roli hrála špatná ekonomická i sociální situace v Čechách (chudoba, vysoké daně, dlouhá vojenská služba), kromě toho byly osadníkům slíbeny dílčí výhody (pozemek na stavbu domu, pole, výjimka na daních i vojenské službě apod.), kterých se jim nakonec nedostalo.

Původ kolonistů není zcela objasněný, ale je zřejmě poměrně heterogenní; obvykle zmiňované jsou Plzeňsko, Klatovsko, Příbramsko, Českobudějovicko a střední Čechy (srov. Salzmann 1984: 70; Urban 1930: 24 aj.).

Osidlování proběhlo ve třech vlnách:

1. 1823–1825: založena Elisabeta (Elisabetfelda, 118)[2] a Svatá Helena (Sfântă Elena, 388)
2. 1826–1830: založen Gerník (Gârnic, Weizenried, 489), Bígr (Bigăr, Schnellersruhe, 266), Eibentál (Eibenthal, 188), Rovensko (Ravensca, 237), Frauewiese (186), Nový Župánek (43), Poňjásky (Schöntal, 281) a Šumice (Şumiţa, 123)
3. po roce 1861: doosidlování Šumice a Rovenska a vesnic se smíšeným obyvatelstvem

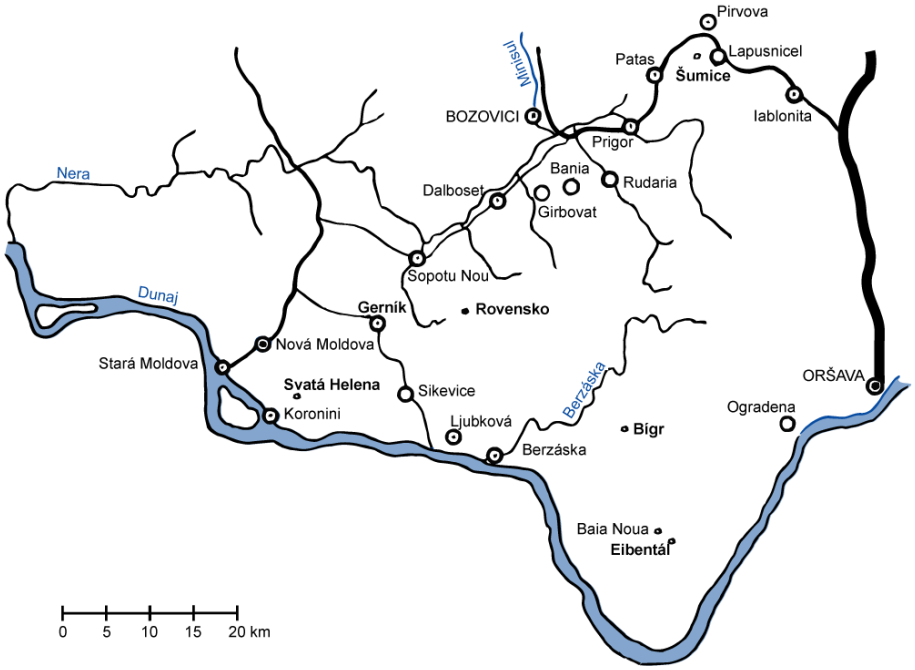
Dodnes se zachovalo šest vesnic s kompaktním českým obyvatelstvem (Bígr, Eibentál, Gerník, Rovensko, Svatá Helena a Šumice; viz mapa 1 na následující straně), zbylé osady zanikly nebo se rumunizovaly. Kromě toho se v Banátu nachází i vesnice a města s poměrně početnou českou menšinou (např. Herkulovy lázně, Klopodie (Klopotín), Nová Ogradena, Oršava, Temešvár, Velký Pereg aj.). Dnes se Eibentál nachází v župě Mehedinti a zbylých pět vesnic v župě Caraş-Severin. V obou celých župách tvoří Češi absolutní většinu obyvatel už více než šedesát let (Costachie et al. 2011: 10).

Od 90. let 19. století se začala populace banátských Čechů zmenšovat kvůli emigraci, a to zejména do rumunských vesnic v okolí, dále pak do Srbska, Bulharska či do Argentiny. Důvodem bylo zejména přelidnění a nedostatek práce.

Po druhé světové válce bylo všem krajanům v zahraničí umožněno přesídlit zpět do Československa (do oblasti Sudet). Této možnosti v letech 1945–1948 využilo zhruba tisíc Čechů z Heleny a Gerníku, celkové odhady uvádějí okolo osmi tisíc Čechů a Slováků z celého Rumunska (Jech et al. 1992: 171). Následné změny poměrů v Československu i v Rumunsku možnost přesídlení na dlouhou dobu uzavřely.

[2] V závorce je uveden v tomto pořadí název rumunský a popřípadě německý společně s počtem obyvatel v roce 1830 dle Urbana (1930: 15). Všechny zeměpisné názvy jsou nadále uváděny pouze v českém úzu zápisu (pokud takový ekvivalent existuje a je běžně užívaný).

V 90. letech došlo ke znovuotevření hranic a kontakty s Českou republikou byly obnoveny. V Banátu od té doby působí projekty české vlády a humanitární organizace Člověk v tísni, o. p. s. Jejich cílem je zpomalit postupný odchod krajanů z banátských vesnic; podporují rozvoj infrastruktury, vzdělání a vznik nových pracovních příležitostí v českých vesnicích. Více než 50 % banátských Čechů odešlo za posledních dvacet let a počet zbývajících obyvatel nadále klesá. Hlavním důvodem odchodu bývá nezaměstnanost a nedostupnost vzdělání i lékařské péče. V dnešní době je ve vesnicích téměř jediným zdrojem příjmů agroturismus, který není dostačující. Vývoj počtu obyvatel shrnuje příloha 2 mojí bakalářské práce (Vyskočilová 2012b).



Mapa 1: Vesnice s kompaktním českým osídlením (autor K. V.)

1.2. Bígr

Z této vesnice pochází prozatím nasbíraný mluvený materiál použitý pro tvorbu korpusu (viz kapitola 5); pro podrobnější informace viz zejména publikace Urban (1930: 23–26), Viková (1994: 4–10), Vyskočilová (2012: 18–19), které slouží zároveň jako zdroj této podkapitoly.

Vesnice byla založena v druhé vlně osídlování v roce 1828. Původ osadníků není zcela jasný, ale rozdílné zdroje se shodují na tom, že sem vystěhovalci přišli z Klatovska, Plzeňska, Příbramska, Chodska a případně dalších krajů (Karas 1937: 47; Urban 1930: 24; Viková

1994: 5). V letech 1848–1850 se do Bígru přestěhovalo i několik českých rodin ze srbského Krušovce (Karas 1937: 44).

Bígr je dodnes jedna z nejodlehlejších vesnic, kterou také navštěvuje poměrně málo turistů. Ročně do Bígru přijede pouze několik stovek turistů a velké množství z nich se nezdrží ve vesnici déle než jednu noc. Srovnávat je možné například s Gerníkem, kam podle slov místního knihovníka, který ubytovává turisty, přijede ročně až 3000 turistů.[3] Ve vesnici se nachází kostel, kulturní dům, dva obchody a jednotřídní škola s osmi postupnými ročníky (viz oddíl 2.1.1).

Bígerští muži původně pracovali několik desetiletí v uhelných a antracitových dolech, ženy zůstávaly doma a staraly se o děti a hospodářství. Dnes je většina obyvatel, která zůstala ve vesnici, v důchodovém věku. Ta část, která neemigrovala do zahraničí, ale přestěhovala se do okolních vesnic či měst, se do Bígru vrací jen na víkendy či na prázdniny.

Dle posledního sčítání lidu žilo v roce 2012 ve vesnici 146 obyvatel (z toho 10 % jiné než české národnosti). Při mé poslední návštěvě v březnu 2014 mi bígerská poštovní doručovatelka sdělila, že obyvatel Bígru není více než sto.

[3] Osobní výzkum na místě, březen 2014.

2. JAZYKY V KONTAKTU – JAZYKOVÁ SITUACE NA ÚZEMÍ BANÁTU

Od osídlení Banátu Čechy v 19. století uplynulo téměř dvě stě let a během té doby se změnila nejen čeština na našem území, ale i čeština banátských krajanů. Velmi dlouho byla tzv. izolovaným jazykem, který se vyvíjel odděleně a nezávisle na svém výchozím jazyku, tj. na češtině (Vašek 1975: 1).

V průběhu odloučení došlo k vyrovnání nářečních rozdílů a uchování hlavních jazykových rysů přinesených z Čech (srov. Salzmann 1983; Skulina 1975, 1978; Utěšený 1962a, 1962b, 1964 aj.). Dochovaly se i dnes již zaniklé anebo čistě nářeční prvky (převážně hláskoslovné a lexikální). Stabilita se nadále projevovala také v zachování antroponym (Alžběta, Bednář, Mleziva aj.) i toponym (Turecká díra) běžných na českém venkově.

Na základě převažujících dialektálních znaků rozdělil Utěšený (1964b: 27–28) banátské vesnice do tří skupin:

1. Bígr a Gerník – západočeské rysy
2. Eibentál, Rovensko a Šumice – středočeské rysy
3. Svatá Helena – středočesko-severovýchodní rysy

Kromě toho docházelo i k ovlivňování cizími jazyky jako je němčina, srbština a rumunština (srov. Lamprecht (1976: 391–392) a Skulina (1975: 69–70, 1978: 157–158)). Zejména vliv rumunštiny je v posledních desetiletích stále intenzivnější.

2.1. Kontakt se spisovnou a obecnou češtinou

Po příchodu do Banátu měli krajané jen velmi omezenou možnost konfrontace s češtinou na našem území, setkávali se s ní pouze prostřednictvím kostela, školy a knih, které si s sebou přinesli. Poté dlouhou dobu kontakt zprostředkovávali čeští učitelé. Až se vznikem Československa se zvýšil zájem Čechů a Slováků o krajany v zahraničí. Začaly se vytvářet krajanské spolky, vydávat časopisy a zakládat knihovny v zahraničí. Po druhé světové válce, kdy část banátských Čechů odešla zpět do vlasti, se krajané v Banátu začali častěji dostávat do kontaktu s obecnou češtinou. V 60. letech následoval útlum vzájemných styků, kontakt byl obnoven až v 90. letech (zejména díky médiím a turismu).

V současnosti přichází krajané do styku se spisovnou češtinou zejména ve školách a při četbě české literatury. Je však vhodné podotknout, že mnohem oblíbenější je sledování

televize, díky níž jsou banátští Češi vystaveni povětšinou obecné češtině (stejně tak v rádiu anebo na internetu).

2.1.1. Školství[4]

Důležitou roli v udržení znalosti českého jazyka hrály školy; první byly vybudovány v 50. letech 19. století v každé vesnici. Čeština se na školách vyučovala vždy, měla ale v rámci vyučování různé postavení. Jedinou výjimkou bylo období let 1907–1918, kdy byly Apponyiho školskými zákony pod tlakem maďarizace zakázány všechny nemaďarské školy (v Banátu uvedeno do praxe až v roce 1910). Na banátských školách působili kromě místních učitelů (rodáků z českých vesnic) i učitelé z Čech, a to do roku 1910 a také v meziválečném období. Změna politického režimu po roce 1948 vedla k přerušení styků mezi Československem a Rumunskem; kontakty byly obnoveny až v 90. letech. Od roku 1998 opět v Banátu působí čeští učitelé, vysílaní Domem zahraniční spolupráce v rámci Programu podpory českého kulturního dědictví v zahraničí.[5] V dnešní době působí v Banátu dva čeští učitelé z České republiky (jeden v Gerníku a Svaté Heleně a druhý v Eibentále).

Úloha českých učitelů nebyla nikdy lehká, často vyučovali češtinu jen několik hodin týdně a navíc v několika vesnicích zároveň. Až do 90. let byl v Bígru vyučován pouze 1. stupeň (1.–4. třída),[6] pro 2. stupeň (6–10. třída, dnes už pouze 6–8. třída) by museli žáci do rumunských škol, což bylo časově i finančně náročné. Do školní docházky se promítaly i sezónní práce na poli, část žáků navštěvovala školu jen v zimě anebo ji vůbec nedokončila. Podle Vikové (1994: 8) vzdělání čtvrtou třídou předčasně ukončilo 80 % děvčat a 10 % chlapců, na středních školách pak pokračovali výhradně chlapci.

Kromě výuky češtiny a předmětů s ní spojených dnes vyslaní učitelé vedou i zájmové kroužky, vyučují angličtinu, často také mají postavení „mediátora“, který zprostředkovává kontakt mezi místními a Českou republikou.

Dlouhou dobu byl česky vyučován celý první ročník, v dnešní době se vyučuje pouze rumunsky. Český jazyk je povinný do čtvrté třídy, od páté do osmé třídy je hlavním jazykem rumunština a z češtiny se stává jen volitelný předmět s dotací tři hodiny týdně. Jedinou

[4] Srov. Gesce (2013: 207–290), Kokaisl (2009: 25–26, 38, 45–46, 53–54) a Moravcová (2006).
[5] *Program podpory českého kulturního dědictví v zahraničí (krajané, lektoři)* [online] [cit. 2014-07-16]. Dostupné z: <http://www.dzs.cz/cz/program-podpory-ceskeho-kulturniho-dedictvi-v-zahranici/>.
[6] Potvrzuje ho ještě Viková (1994: 7).

výjimkou je Eibentál, kde je čeština volitelným předmětem již od první třídy (Costachie et al. 2011: 11).

Školy v jednotlivých vesnicích se liší nejen vybavením učeben, ale výrazně také počtem žáků, což dále ovlivňuje způsob výuky. Jako příklad lze uvést školu v Bígru, která je jednotřídní a v březnu 2014 ji navštěvovali pouze čtyři žáci. Podle slov místního učitele Mlezivy škole hrozí zavření již v příštím roce, kdy se jejich počet sníží na pouhé dva žáky (MŠ by zavřena už v 90. letech). Opakem je situace v Eibentálu, který má školu s největším počtem žáků: 10 dětí v MŠ, 31 žáků v ZŠ (ve školním roce 2012/13). Žáci jsou vyučováni po jednotlivých třídách tak, jak je tomu i u nás (Roháľová 2013).[7]

2.1.2. Média a komunikace

Banátští Češi mají možnost sledovat českou televizi (zejména ČT24) a pořady v češtině (tři rumunské kanály vysílají i české seriály a pořady v češtině s rumunskými titulky). Televizi se satelitem má doma téměř každý a jedná se tak o nejfrekventovanější sdělovací prostředek, se kterým banátští Češi přijdou do styku. Kromě toho je možné naladit Rádio Temešvár a Rádio Rešice, která vysílají půlhodinové pořady speciálně vytvořené pro českou komunitu (Lazu 2010).

Co se psaných médií týká, noviny a časopisy si lze předplatit pouze v rumunštině. Možnost přístupu na internet se liší od vesnice k vesnici. V Bígru je poměrně kvalitní internet ve škole, čehož místní využívají zejména pro komunikaci s příbuznými v Čechách přes službu Skype. Od letošního roku mají místní možnost si zavést přípojku i domů (což ale ještě v březnu 2014 nikdo neudělal). Stejná situace je i v Gerníku a domácnosti, které pravidelně ubytovávají turisty, mají internet dokonce doma a aktivně ho využívají ke komunikaci s nimi. Oproti tomu v Eibentálu je pouze velmi špatný internet, který nepostačuje ani učitelům.

2.2. Kontakt s cizími jazyky

Jak už bylo řečeno výše, Češi v Banátu se po celou svou historii dostávali do kontaktu s cizími jazyky. Nejdříve němčinou, která byla oficiálním jazykem celé Habsburské

[7] Počet žáků v ostatních vesnicích je podobný (stav ve školním roce 2012/13): Gerník – 2 MŠ, 14 ZŠ; Svatá Helena – 10 MŠ, 29 ZŠ (Skořepa 2013a, 2013b). Počet žáků na Rovensku a v Šumici se mi nepodařilo zjistit.

monarchie. Nikdy ale nedošlo k vyvinutí česko-německého bilingvismu. V dnešní banátské češtině jsou přesto stále používaná slova jako *cuk* (vlak), *špolhert/šporhet* (kamna) aj. Po připadnutí Banátu Uhrám v roce 1868 roli oficiálního jazyka převzala maďarština. Vliv maďarštiny se ale na banátské češtině téměř neprojevil, používaný zůstal jen např. název vesnice *Ujbánie* (rum. *Baia Nouă*).

Rozdělení historického Banátu po první světové válce přineslo jako oficiální jazyk rumunštinu, což se projevilo i v českých vesnicích. Přestože rumunská vláda ve školách povolila vyučování v českém jazyce, nedostatek českých učitelů vedl k tomu, že se dlouho učilo jen rumunsky (pro více informací viz oddíl 2.1.1). Od té doby se vliv rumunštiny na češtinu neustále zesiluje, roli hraje zejména otvírání vesnic okolnímu světu. Už Skulina (1975: 69) poznamenává, že „česká i slovenská menšina se podle soustředěnosti sídel i podle sociálního složení asimiluje s rumunským jazykovým prostředím“. Interference mezi češtinou a rumunštinou vede k vzájemnému asimilačnímu působení a stejný vliv může mít na banátskou češtinu i bilingvismus českých mluvčích.

Kromě těchto jazyků se projevuje také vliv srbochorvatštiny, která přinesla i výpůjčky z bulharštiny a turečtiny. Ne vždy je původ vypůjčených prvků zřejmý, protože v oblasti žijí nebo žily četné národnostní menšiny a navíc je část slov společná všem těmto jazykům. Vypůjčeným výrazům se říká banatismy (viz Skulina 1975: 70).

Pokud jde o fonetickou a morfologickou stránku vypůjčených výrazů, všechna slova cizího původu se přizpůsobila potřebám češtiny (Ciplea 1971: 215). Lexikálním výpůjčkám do banátské češtiny se věnuje velké množství studií, např. Ciplea (1966, 1968, 1971), Dobriţoiu-Alexandru (1967), Skulina (1975, 1978) a Utěšený (1962, 1964b). Interference se samozřejmě v menší míře projevuje i v morfologii, syntaxi či frazeologických kalcích (srov. předchozí zmiňované a Vašek 1968).

2.2.1. Bilingvismus

Na území Banátu historicky vždy panovala značná diglosie. Na otázku, od kdy jsou banátští Češi bilingvní, neexistuje jednoznačná odpověď, ale vznik bilingvismu je patrně možno datovat až od 70. let 20. století. Dobriţoiu-Alexandru (1967: 375) ve své studii píše, že „[…] ve zmíněných banátských obcích s téměř výlučně českým obyvatelstvem není bilingvismus masovým jevem“ a že je přítomný pouze v Klopodii. Tam bilingvismus českých mluvčích potvrzuje Vašek (1975). Ve stejné době píše o bilingvismu mladší generace

v Šumici Skulina (1975: 70), v případě starších krajanů zmiňuje bilingvismus jen u těch, kteří žili nějakou dobu mimo vesnici. V pozdějších článcích už Skulina (1978: 162, 1979: 227) zmiňuje bilingvismus bez výhrad. Viková (1994: 10) o patnáct let později bilingvismus potvrzuje s tím, že mladší generace je jím ovlivněna více.

Jak jsem shrnula ve svém článku (Vyskočilová, v tisku), dnešní česká menšina v Banátu tvoří podle mého mínění poměrně jednotnou řečovou komunitu. Její mluvčí jsou zvyklí volit mezi češtinou a rumunštinou podle toho, jaký jazyk komunikační situace vyžaduje. Znalosti rumunštiny je však stále spíše pasivní – slouží ke zjišťování informací ze sdělovacích prostředků. Zřídka čeští mluvčí využijí rumunštinu v komunikaci s Nečechy anebo s mluvčími, kteří preferují komunikovat pouze rumunsky, a to buď ze zdvořilosti anebo ve snaze ukázat svou znalost jazyka. Obecně řečeno, banátští Češi rozumí informacím v rumunštině, ale jejich schopnost používat rumunštinu aktivně záleží na mnoha individuálních faktorech: talent, věk, pohlaví, vzdělání, zaměstnání aj. Zejména mladí lidé a děti přepínají mezi jazyky téměř automaticky. K bilingvismu mluvčích dále patří to, že jsou schopni přizpůsobit svůj projev posluchačům, a to i v případě setkání s mluvčími obecné češtiny, protože si uvědomují jazykové odlišnosti.

V českých banátských vesnicích kromě Čechů žije i malé procento osob jiných národností (dle posledního sčítání lidu to bylo v roce 2011 5–10 % v závislosti na vesnici). V případě, že jsou tyto osoby do komunity integrovány, často se na různé úrovni naučí česky a češtinu používají v komunikaci s krajany.

3. LINGVISTICKÝ VÝZKUM BANÁTSKÉ ČEŠTINY

Přestože první zmínky o jazyce banátských Čechů pocházejí ze začátku 20. století, první skutečně lingvistický výzkum je datován až do 60. let 20. století. Výzkum banátské češtiny měl dvě výrazná období, a to především 60. a 70. léta a pak období od počátku 21. století až do současnosti. Kromě samotného výzkumu je pro tato období mimo jiné charakteristické i to, že zatímco výzkumníci z prvního období byli již „zavedení" lingvisté, období druhé patří studentům a jejich závěrečným pracím.

První lingvistické poznatky přinesly dva terénní výzkumy Ústavu pro etnografii a folkloristiku, kterých se zúčastnil i dialektolog Slavomír Utěšený. Nasbíral jazykový materiál v Bígru, Eibentálu, Gerníku a Svaté Heleně a vydal několik studií (Utěšený 1962, 1964a, 1964b, 1970), ve kterých se věnoval základům všech jazykových plánů, zejména lexikologii a morfologii. Jeho výzkum v Gerníku, Svaté Heleně a Velkém Peregu byl zařazen do *Českého jazykového atlasu* (ČJA) a Gerník do *Českých nářečních textů* (Lamprecht 1976). Jako doplněk ČJA později vyšlo CD *Jak se mluvilo v českých vesnicích v cizině* (Bachmannová a Jančák 2002), které zahrnuje Utěšeného ukázku mluvy z Gerníku.

V 70. letech se zabývá banátskou češtinou v Šumici Josef Skulina (1974, 1975, 1978, 1979), který se zaměřuje taktéž na morfologii a lexikologii, a dále Antonín Vašek (1968), zkoumající jazyky v kontaktu, mimo jiné na materiálu z rumunské Klopodie, kde žila také česká menšina. Výsledky obou byly taktéž zařazeny do ČJA a *Českých nářečních textů* (Lamprecht 1976).

Ve stejné době se k banátské češtině obrací i pozornost rumunských bohemistů – Gheorghe Ciplea (1966, 1968, 1971), Tiberiu Pleter (1965) a Theodora Dobriţoiu-Alexandru (1965, 1967). Ve svých studiích se věnovali převážně vlivu rumunštiny na fonetiku, morfologii a zejména slovní zásobu, v případě poslední zmíněné autorky i historii českých osad.

První období uzavírá lingvistický antropolog Zdeněk Salzmann (1983, 1984, 1993). Věnoval se ponejvíce jazyku a kultuře v Šumici. Kromě toho pořídil detailní soupis bibliografie do roku 2004 (Salzmann 2004) o celé široké oblasti výzkumu věnované krajanům v Rumunsku (etnologie, folkloristika, historie, jazykověda, sociologie aj.).

3.1. Stav od 90. let 20. století do současnosti

Jak bylo poukázáno výše, v posledních dvaceti letech se na Banát zaměřují zejména práce studentské (bakalářské a diplomové). Jedná se stovky děl studentů z různých oborů, z dostupných repozitářů závěrečných prací se přitom zdaleka nedají získat informace o všech. Rozpětí oborů, ze kterých přicházejí studenti věnující se banátské problematice, je velice široké, vzhledem k zaměření této práce zde tudíž zmíním pouze studie lingvistické, kterých zdaleka není takové množství.

Krátce po otevření hranic se české menšině věnuje Vilma Viková (1994) ve své diplomové práci *Bígerská čeština* s podtitulem *Nástin jazykové monografie české vesnice v Rumunsku*. Tato práce analyzuje jazykově odlišné jevy bígerské češtiny oproti češtině na našem území, které se zachovaly ze středočeských a jihozápadočeských nářečí 19. století. Výzkum probíhá na všech jazykových rovinách (jako jedna z mála se Viková věnuje podrobně i syntaxi), a to na základě srovnání nahraného materiálu a spisovné češtiny. V závěru práce je přiložen slovník nejfrekventovanějších diferenčních výrazů banátské češtiny. Viková shrnuje, že v bígerské češtině došlo k vyrovnání rozdílů mezi nářečími, která si s sebou banátští Češi přinesli ze své vlasti. Bígerskou češtinu již za nářečí českého jazyka nepovažuje, protože v ní dochází k odlišnému vývoji oproti češtině na našem území – tedy zejména k ovlivňování rumunštinou.

Existují záznamy (mimo jiné na ně odkazuje např. Haiderová (2007)), že na Univerzitě Palackého v Olomouci vznikla diplomová práce *Čeština Čechů v rumunském Banátě – Rovensko* P. Honové (2000), kterou se mi ale bohužel nepodařilo získat. Navíc na Katedře českého jazyka a literatury UP existenci práce popřeli, ač koli figuruje v jejím knihovním systému.

Podobnou strukturu jako výše zmíněná Viková zvolila i Karolina Haiderová (2007) ve své diplomové práci *Jazyk české menšiny v rumunském Banátu: obce Gerník a Svatá Helena*. Zabývá se současným stavem jazyka (v promluvách mladé, střední a staré generace) s ohledem na zachovalost původních nářečí. Haiderová využívá dotazníkového šetření, které doplňuje rozborem zvukových nahrávek. Výzkum pojímá diachronně a výsledky srovnává se stavem, který popsal Utěšený ve výše zmíněných studiích. Dochází k závěru, že archaičtější rysy jsou lépe uchovány v gernické češtině, a to i v mluvě mladé generace, zatímco ve Svaté Heleně ustoupily spisovným podobám.

Na svou práci později Haiderová navázala ještě třemi přednáškami na konferencích, ze kterých vyšly články „Archaismy v mluvě české menšiny v rumunském Banátu“ (Haiderová 2008), „Jazyk mladé generace rumunských Čechů v Banátu“ (Haiderová 2010a) a „K vývojovým tendencím v deklinaci substantiv v současné rumunské češtině“ (Haiderová 2010b).

Svým přístupem je ojedinělá bakalářská práce Markéty Tůmové (2011) *Jazyk rumunských Čechů reemigrantů – současný stav*, protože se jako jediná zabývá jazykem reemigrantů. Ke zvolenému tématu se staví převážně ze sociolingvistického hlediska. Materiálem jsou polořízené rozhovory se třemi Čechy narozenými v Banátu, kteří reemigrovali téměř před dvaceti lety. Autorka zkoumá proces jazykové adaptace a asimilace banátských Čechů, ověřuje, jak vnímají obě jazykové varianty, a zjišťuje, jakým způsobem se s češtinou na našem území seznamovali před emigrací. Analýza je navíc doplněná o stručný rozbor promluv z hlediska fonetiky, lexikologie, morfologie a syntaxe na základě výsledků studie Haiderové (viz výše). Podle Tůmové v jazyce reemigrantů přetrvávají jen ty znaky, které jsou zároveň používané i v českých nářečích.

V roce 2012 vznikly hned dvě práce o jazyce v Banátu. Autorkou první z nich je Adéla Frnochová (2012), která napsala diplomovou práci s názvem *Jazyk české menšiny v obci Šumice v rumunském Banátě*. Vychází ze Skulinových článků (viz výše) a pro popis soudobé šumické češtiny na všech jazykových rovinách jí slouží dotazníkové šetření a několik hodin nahrávek mluvených projevů (spontánní rozhovory i vyprávění). Srovnávacím materiálem jsou přepisy nahrávek z roku 1971 pořízené Skulinou a lexikální doklady ze Šumice zveřejněné v ČJA. Zvláštní pozornost byla věnována nářečním rysům (na rozdíl od Skuliny charakterizuje místní nářečí jako jihozápadočeské, nikoliv středočeské, a to na základě slovních forem vyskytujících se pouze v dané oblasti), archaickým jevům a ovlivnění rumunštinou a němčinou.

Druhou prací je bakalářská práce autorky tohoto textu, *Syntaktická analýza projevů českých mluvčích v rumunském Banátu* (Vyskočilová 2012b), která na základě vytvořeného korpusu mluvené češtiny BANÁT rozebírá jednotlivé syntaktické jevy, zmíněné v předchozích výzkumech a porovnává stav bígerské češtiny se stavem obecné češtiny zachyceným v korpusu ORAL2006 a ORAL2008. Na tuto práci autorka navázala v článcích „Korpus BANÁT“ (Vyskočilová 2012a) a „K slovosledu mluvené češtiny v rumunském Bígru“ (Vyskočilová 2013) a v neposlední řadě také touto diplomovou prací. Příští rok by měl

být publikován sociolingvisticky zaměřený článek „Czech language minority in the South Western Romanian Banat“ (Vyskočilová, v tisku).

Poslední prací, která je v době vzniku tohoto textu známá, jsou *Vybrané kapitoly z onomastiky a dialektologie vesnice Gerník (rumunský Banát)* Heleny Orálkové (2013). Tato diplomová práce se věnuje dvěma hlavním tématům. Prvním je zkoumání antroponym (rodná jména, příjmení, přezdívky, hypokoristika a jména po chalupě) a toponym na základě dotazníku. V druhé části jsou na základě ústního pohovoru se čtyřmi informátory (střední a starší generace) ověřována hesla z ČJA 1–3 vlastní Gerníku a porovnávána se současnou tamní lexikální slovní zásobou. Autorka nepotvrzuje žádnou výraznou změnu v tomto typu slovní zásoby, kromě toho, že se objevily i jiné tvary vybraných slov, které pro tuto nářeční oblast nejsou v ČJA uvedeny. Je ale třeba poznamenat, že se výzkum bohužel omezil jen na verifikaci starých slov a nezaměřil se už na zkoumání slov nově vzniklých či přejatých za poslední půlstoletí.

4. SPECIALIZOVANÉ KORPUSY

Tato kapitola rozebírá a shrnuje problematiku specializovaných korpusů, a to zejména s ohledem na specializované mluvené korpusy českého jazyka, mezi které vytvářený korpus banátské češtiny patří.

S ohledem na to, že je kapitola z části vystavěna na anglických zdrojích, bylo třeba vyřešit otázku překladu termínů, který bohužel v oboru není vždy jednotný. Mnohdy se česká terminologie teprve vytváří, tudíž je vhodné uvést již zavedený anglický ekvivalent, jindy naopak existuje jeden český zavedený termín pro celou škálu termínů anglických. České termíny používané v této kapitole a kapitolách následujících vychází zejména z citovaných studií, pro přehlednost a případně i jednoznačnost jsou nicméně v závorce doplněny anglickým termínem.

4.1. K definici specializovaných korpusů

Podle šíře popisu jazyka můžeme dělit jazykové korpusy podle homogenity (*internal consistency*) na obecné (obvykle jen *corpus*) a specializované (*special/specialized corpus*). Obecný korpus se snaží zachytit jazyk celkově či z velké části, je heterogenní a za ideálních podmínek se snaží přiblížit kritériím vyváženosti (*balance*), reprezentativnosti (*representativeness*) a srovnatelnosti (*comparability*) (viz Atkins et al. 1991; Biber 1993). V opozici k němu stojí korpus specializovaný, který se zabývá jen úzce vymezenou oblastí (např. korpus politických debat), je obvykle homogenní a hledisko reprezentativnosti či vyváženosti bývá zjednodušené (Křen 2013: 18–20).

Starší a přesto často citovanou definici specializovaného korpusu nabízí Sinclaire (1996: 6): „Od korpusu se očekává, že bude mít určité vlastnosti spojené s výchozími hodnotami. Není-li uvedeno jinak, tyto vlastnosti jsou připisovány všemu, co se nazývá korpus. Korpus, který má jednu nebo více hodnot jiných než jsou hodnoty výchozí pro danou vlastnost je nazýván korpusem specializovaným: jeho název by měl přibližovat odchylky od očekávaných hodnot.“[8] V další části Sinclaire (1996: 7) dodává, že „pokud lingvista při

[8] „A corpus is assumed to have certain characteristics attached with default values. Unless stated these characteristics are attributed to anything called a corpus. A corpus which has one or more nondefault values

tvorbě korpusu překročí hranice soudržnosti, která je nezbytná pro získání dat, tak by měl přistoupit k vytvoření specializovaého korpusu.“[9]

Problém tkví v tom, že Sinclairem zmiňované vlastnosti obecných korpusů jsou následující: velký počet slov v korpusu; autentičnost v podobě přirozených rozhovorů na běžná témata; jednoduchost při zpracování v podobě textu, který je možné dále jednoduše zpracovávat, ale neobsahuje další lingvistické informace týkající se textu; dobrá dokumentace (Sinclaire 1996: 6–8). V dnešní době už jsou některé parametry minimálně z části pozměněny; specializované korpusy jsou sice oproti obecným korpusům relativně malé, protože sběr dat je obvykle mnohem složitější, ale i tak je možné vytvořit velký korpus, který je možné charakterizovat jako speciální, např. v oblasti mediální komunikace. Otázka autentičnosti je sporná i v případě korpusů obecných; Sinclairova (1996: 7) definice autetičnosti je následující: „Všechna data jsme získali ze skutečných konverzací, které lidé vedou v každodenním životě.“[10] To jinými slovy znamená, že text musel vzniknout jen kvůli komunikaci bez ohledu na jeho další vědecké použití (za autentické je možno považovat i zásahy cenzora nebo redaktora, i když změní původní text bez ohledu na záměr autora). Obzvláště u korpusů psaného jazyka není snadné rozlišit autentický a neautentický psaný text. Další lingvistické informace či nutná metadata k textu jsou běžnou součástí korpusů a naopak jejich absence (např. u webových/internetových korpusů (*internet/web corpus*)) může být problematická pro interpretaci nasbíraných dat.

Vzhledem k tomu, že patrně nelze vyjmenovat budoucí možné způsoby korpusové specializace, se domnívám, že je třeba pojetí definice specializovaných korpusů v dnešní době aktualizovat a uvolnit a zcela obecně je pojmout jako korpusy zaměřené na sledování určité oblasti jako primárního cíle. Ačkoliv zcela jistě existují prototypické příklady pro korpusy obecné (v českém prostředí např. korpusy řady SYN ÚČJTK) i specializované (např. akviziční korpusy AKCES), je třeba mít na paměti, že hranice mezi obecným a specializovaným korpusem není zcela jasná a pevně daná.

---

for these characteristics is termed a special corpus: its title should specify its deviations from the assumptions.“ Všechny překlady použité v této práci jsou dílem autorky.

[9] „[a]nything which involves the linguist beyond the minimum disruption required to acquire the data is reason for declaring a special corpus.“

[10] „All the material is gathered from the genuine communications of people going about their normal business.“

4.2. K tvorbě korpusů[11]

Při tvorbě obecného korpusu je v dnešní době cílem co nejvíce se přiblížit ideálu vyváženosti, reprezentativnosti a srovnatelnosti, tak aby byl výsledný podíl stejný u všech zastoupených jazykových variet (*domain*, *register*, *sublangue*, *text type*, *genre*, *topic*, *style*), do kterých je třeba nasbíraná data zařadit podle určitých kritérií. Jak bylo zmíněno v předchozí podkapitole, větší či menší část těchto hodnot může být ve specializovaných korpusech ovlivněna primárním zaměřením korpusu. Důvody jsou často nelingvistické – když se jedná např. o jazyk neuznávané skupiny, tak ani výzkum jazyka není podporován, u ohrožených jazyků je tento fakt naopak dán tím, že existuje pouze málo mluvčích.

Významný problém potom představuje zpracování mluvených korpusů takových jazyků, protože je časově i finančně náročné a bez náležité podpory nemožné zpracovat velké množství dat. Tvůrce korpusu je často veden pragmatismem a pracuje s objemem dat odpovídajícím jeho záměru. „I přes velké množství práce a příprav se může ukázat, že sestavit ideální korpus pro ten který jazyk je nemožné. Pokud se jedná o zanikly nebo zanikající jazyk a data potřebná k vytvoření velkého, vyváženého korpusu nejsou a nikdy nebudou k dispozici.“[12] (McEnery a Hardie 2012: 12). Stručně řečeno, je třeba přijmout skutečnost, že mohou a musí být vytvářeny i takové korpusy, které jsou formovány pragmatickými rozhodnutími (McEnery a Hardie 2012: 11–13).

Křen (2013: 18) komentuje situaci mluveného korpusu ORAL 2008, který je z hlediska reprezentativnosti sociolingvisticky vyvážený dle pohlaví, věku, vzdělání a místa pobytu v dětství. Křen nicméně tato kritéria nepokládá za dostačující, protože je velmi zjednodušená, a řadí ho proto mezi specializované, homogenní korpusy.

Pro kategorizaci dat se používají dva odlišné přístupy, které jsou založené buď na interních, nebo na externích kritériích (*internal and external criteria*). Externí kritéria jsou mimotextové rysy, které se vztahují k funkci textu s ohledem na účel jeho produkce, a

[11] Srov. Atkins et al. (1991); Biber (1993); Křen (2013: 16–22), McEnery a Hardie (2012: 11–13); Sinclair (1996).

[12] „Finally, even with a huge amount of work and planning, it may simply be impossible to build an ideal corpus for a given language – if the language is dead or dying and the material to construct a large, balanced corpus is not available and simply never will be.“

obvykle jsou předem daná. Interní kritéria jsou vystavěná pouze na základě jevů přítomných v textu bez ohledu na kritéria externí (nejrůznější lexikální a gramatické vztahy).

Vzhledem k tomu, že se s externími kritérii lépe pracuje, protože jsou již uživatelům známá (např. dělení textů na publicistiku, beletrii apod.), uplatňují se při tvorbě korpusů častěji. Ale je to právě intuitivnost těchto kritérií, která se může stát problémem, protože dělení může být značně subjektivní a není jednoduché být zcela konzistentní. Interní kategorie jsou sice snadno kvantifikovatelné, ale špatně uchopitelné v rámci komunikačních situací i v otázce přístupu k vyváženosti takového korpusu. Interní kritéria není možné vytvořit před sběrem samotného materiálu, proto se ponejvíce používají při získávání zpětné vazby ok nasbíraným datům.

Vhodná velikost korpusu je často diskutovaným problémem, na který neexistuje jednoznačná odpověď, protože ta souvisí se sledovanými jevy. Obecně platí, že pro frekventované jevy (např. jevy gramatické) stačí menší korpusy, ale pro méně časté jevy (např. jevy lexikální) je třeba korpusů větších. V případě lexikologie je možné softwarově[13] naplánovat velikost korpusu tak, aby zahrnoval počet vybraných slov s požadovanou frekvencí (Milička 2012). Nicméně ve většině ostatních případů není možno jednoduše vyjádřit, co je velký, malý či dostatečný korpus. Podle Křena (2013: 20) historicky pořád platí, „že maximální velikost korpusu je dostatečná, ať už jí byl 1 milion slov korpusu Brown na začátku 60. let, 100 milionů BNV o třicet let později nebo současné miliardy ve webových korpusech, IDS Mannheim nebo ČNK“. Přesto zápory nevhodně složeného korpusu nemohou být vyváženy jeho velikostí.

V mluvených korpusech je často diskutovaným problémem spontánnost projevů a názory lingvistů se často radikálně liší. Podle některých je pro dosažení spontaneity nutné jít až na hranici etiky a pořídit záznam bez vědomí mluvčího s jeho dodatečným souhlasem (to mj. předpokládá dobrý vztah mezi nahrávajícím a respondentem, což umožní odhadnout očekávané „ano“ již před samotným nahráváním). Tento postup je např. doporučován pro sběr materiálů pro korpus ORAL2013 (Benešová a Waclawičová 2014). Protipólem je tvrzení, že i

[13] MILIČKA, Jiří, nedatováno. *Lexicographers' Calculator* [online] [cit. 2014-08-13]. Dostupné z: <http://milicka.cz/lexicographerscalculator/>.

v technicky manipulovaných a akusticky ošetřených situacích, může vzniknout spontánní projev, neboť jsou tyto faktory často kompenzovány pomocí hovoru s důvěrně známou a blízkou osobou (Křivan a Zíková 2014: 81).

Křivan a Ziková (1994: 81) dále shrnují známý fakt, že při přirozeném projevu, který se odehrává v mluvčímu známém prostředí, je menší šance, že se nahrávajícímu podaří zajistit ideální podmínky pro nahrávání a tím kvalitní výstup (obzvláště pro fonetickou analýzu). Nahrávaný totiž o nahrávání v tomto případě neví a mikrofon je někde skrytý, a to jen těžko na vhodném místě a zároveň za zcela ideálních okolností (žádné ruchy nebo odrazy, dostatečná blízkost k mluvčím apod.).

4.3. Typy specializovaných korpusů

Tradičním a jasně vymezeným příkladem specializovaných korpusů jsou korpusy akviziční (*acquisition corpus*), tj. speciální korpusy vytvářené za účelem studia osvojování jazyka a jeho vyučování (srov. Bedřichová et al. 2011; Behrens 2008; Granger 2002; Šebesta 2010; Šebesta a Škodová 2012). Využití najdou nejen při samotném studiu akvizičních procesů, ale také v pedagogické oblasti, kde mohou být využity např. přímo ve výuce nebo při přípravě studijních materiálů (učebnice, slovníky, testy aj.).

Pod tento název se obvykle zahrnují dva typy korpusů podle svého zaměření. Prvním z nich je korpus jazyka dětí/mládeže (*developmental corpus*), mapující proces osvojování prvního (mateřského) jazyka. Druhým typem jsou korpusy žákovské/studijní (*learner corpus*), věnující se procesu akvizice druhého/cizího jazyka. Podle způsobu sběru jazykového materiálu se akviziční korpusy dělí na korpusy průřezové/příčné/transverzální (*synchronic corpus*), zachycující jazyk za jedno časové období, longitudiální (*longitudial corpus*), sledující jazyk stejných žáků v několika různých obdobích, pseudolongitudiální (*quasilongitudial corpus*), homogenní skupina žáků s rozdílnými znalostmi v jednom časovém období; případně dále korpusy smíšené. Dostupné korpusy jsou z velké části průřezové či pseudolongitudiální z toho prostého důvodu, že vybudování longitudiálního korpusu je velice náročné, protože vybranou populaci je třeba sledovat měsíce či roky.

V metadatech k akvizičním korpusům jsou standardně zaznamenávány jak údaje o mluvčím, tak i podmínky a způsoby získávání materiálu. Pro akviziční korpusy je také charakteristická často používaná chybová anotace, tj. vyznačení odchylky od jazykové normy.

Pokud je mi známo, další typy specializovaných korpusů nejsou v odborné literatuře tematizovány. Domnívám se však, že je možné takové skupiny definovat na základě společných témat či charakteristik. Speciální korpusy jsem s ohledem na jejich zaměření vedle korpusů akvizičních rozdělila na tři další skupiny: (1) korpusy dialektální či korpusy zachycující vybranou varietu jazyka; (2) korpusy mediálních promluv a komunikace; (3) korpusy fonetické. Toto dělení na jednotlivé skupiny bylo vytvořeno zejména za účelem utřídění informací o existujících specializovaných mluvených korpusech českého jazyka v následující podkapitole.

Nejvolnější (a patrně původně zcela splývající s označením specializovaný korpus) je kategorie korpusů dialektálních či korpusů zachycujících vybranou varietu jazyka. Dá se předpokládat, že z tohoto typu se budou s přibývajícími korpusy v budoucnu postupně vydělovat další typy (stejně jako tomu bylo v případě vydělení akvizičních a mediálních korpusů). Jejich společným jmenovatelem je nejčastěji stejný původ/situace mluvčích či shodná lokalita nahrávání.

Korpusy mediálních promluv a komunikace (či zjednodušeně korpusy publicistické) by mohly patřit do předchozí skupiny korpusů, ale vzhledem k jejich množství a stejnému způsobu nahrávání nebo získávání nahrávek (kdy velice často stačí zpracovat již dostupné nahrávky a není třeba je sbírat) si myslím, že si zaslouží svoji vlastní skupinu stejně jako korpusy akviziční.

Korpusy, které označuji jako korpusy fonetické, jsou budovány především se záměrem zkoumat fonetické jevy na daném materiálu, často by ale mohly být řazeny i do dalších zmíněných kategorií. Fonetický výzkum se vyznačuje zejména velkými nároky na kvalitu nahrávek a často velmi specifickým a detailním způsobem přepisu takového materiálu, což považuji za základní rozlišovací kritérium. Nárokům na kvalitu nahrávek se musí často podřídit autentičnost nahrávek, protože ideální prostředí pro nahrávání ne vždy bývá přirozeným prostředím pro produkci daných textů (k nahrávání viz např. Křivan a Zíková 2014).

4.4. Specializované mluvené korpusy českého jazyka

V této podkapitole je mojí snahou vytvořit základní přehled vybraných existujících či právě vznikajících specializovaných mluvených korpusů českého jazyka. Jejich soupis zdaleka není úplný, a to proto, že se jedná obvykle o malé korpusy vzniklé pro potřeby

jednotlivých pracovišť a často bývají spíše v pozadí jednotlivých výzkumů. Nicméně pokládám za užitečné se o takový přehled alespoň pokusit, protože žádný podobný seznam neexistuje,[14] přičemž ale lze na konkrétních korpusech zároveň ukázat, jakým způsobem může být korpus specializovaný.

4.4.1. Korpusy akviziční a školské komunikace

Mezi české mluvené akviziční korpusy bezpochyby patří dva korpusy řady AKCES[15] budované na Ústavu českého jazyka a teorie komunikace (ÚČJTK) (srov. Bedřichová et al. 2011; Šebesta 2010; Šebesta a Škodová 2012). Patří sem právě dokončovaný korpus mluvených projevů romských dětí a mládeže ROMi 1.0,[16] který má rozsah 120 000 slov získaných od 143 mluvčích věkem mezi 13 a 24 lety a je nahraný v prostředí škol a neziskových organizací často formou řízeného rozhovoru (set otázek či témat). Zcela unikátní bude longitudiální korpus CZESL-LONG/CZEFL-LONG (Czech as a Second/Foreign Language), s jehož sběrem se právě začíná. Mělo by se jednat o longitudinální korpus mluvených projevů žáků, založený na čtyřletých sběrech probíhajících několikrát za rok na vybraných gymnáziích, středních školách a na univerzitě v Káhiře. Korpus by měl sestávat z nahrávek vyučovacích hodin a diskuse několika žáků ve skupinách s nahrávajícím.

O poznání větší je Korpus vyučovacích hodin SCHOLA2010,[17] vzniklý na ÚČJTK a zpřístupněný v nabídce korpusů Českého národního korpusu (ČNK). Jedná se o korpus zaznamenávající komunikaci žáků i studentů během vyučovacích hodin. Celkem bylo během let 2005–2008 pořízeno 204 nahrávek o celkové délce 143 hodin 25 minut a rozsahu téměř 800 000 slov. Nahrávky byly pořízeny ve všech ročnících[18] na 27 základních školách, gymnáziích a odborných školách a účastnilo se jich 2347 žáků ve věku 6–23 let a 47 pedagogů ve věku 23–53 let. Vyvážený je tento korpus s ohledem na skupiny vyučovacích předmětu (český jazyk a literatura; matematické a přírodovědné předměty; společenskovědní a výchovné předměty; informatika a technické a profesně-pracovní vyučování.

---

[14] Až těsně před tiskem této diplomové práce se mi dostalo informace o novém článku (Kopřivová et al., v tisku), který se zčásti věnuje podobnému přehledu, proto na něj odkazuji pouze zde.
[15] *Akces: Akviziční korpusy českého jazyka* [online] [cit. 2014-08-07]. Dostupné z: <http://akces.ff.cuni.cz/>.
[16] *ROMi 1.0* [online] [cit. 2014-08-07]. Dostupné z: <http://ufal.mff.cuni.cz/romi-10>.
[17] GOLÁŇOVÁ, Hana a Karel ŠEBESTA, nedatováno. *Korpus vyučovacích hodin SCHOLA2010* [online] [cit. 2014-08-07]. Dostupné z: <http://ucnk.ff.cuni.cz/schola.php>.
[18] S výjimkou 4. třídy základní školy, kde se nahrávání nepodařilo zajistit.

Stejný kolektiv by měl pracovat v budoucnu i na Korpusu neformální komunikace dětí a mládeže a Korpusu dětí a mládeže v mediální komunikaci (Goláňová a Matějů 2008). Ani k jednomu se mi bohužel nepodařilo získat bližší informace.

4.4.2. Korpusy dialektální či zachycující vybranou varietu jazyka

Do této kategorie patří poměrně dobře známé a hojně využívané korpusy mluveného jazyka ČNK, z toho důvodu se neuchyluji k podrobnějšímu popisu. Jedná se jmenovitě o korpusy neformální mluvené češtiny ORAL2006, ORAL2008 a o zvukovou stopu doplněný ORAL2013 (Benešová et al. 2013; Waclawičová et al. nedatováno; Waclawičová a Richterová nedatováno) a také o Pražský mluvený korpus (PMK) a Brněnský mluvený korpus (BMK) s promluvami z 90. let 20. století (Čermák nedatováno; Hladká nedatováno). Dále sem lze zařadit podobně koncipovaný Olomoucký korpus mluvené češtiny (OMK), budovaný od roku 2002 na Filozofické fakultě v Olomouci (Pořízka 2007).

Podrobnější informace o zmiňovaných korpusech jsou shrnuty v následující tabulce:

		počet hodin	počet tokenů	datum sběru	místo sběru
BMK		55	596 009	1994–1999	Brno
OMK	**OL**	140	1 500 000	od 2002	Olomouc
	CZ	220	2 000 000	od 2010	celá ČR
ORAL2006		112	1 312 282	2002–2006	Čechy
ORAL2008		115	1 349 536	2002–2007	Čechy
ORAL2013		300	3 285 508	2008–2011	celá ČR
PMK		100	819 267	1988–1996	Praha a okolí

Tabulka 1: Přehled dialektálních mluvených korpusů

Dále sem patří i projekt výstavby diachronního korpusu DIALEKT, který by měl zpřístupnit data z 60. a 70. let 20. století z nářečních oblastí na území České republiky (Kopřivová et al. 2014).

Kromě výše zmiňovaných relativně velkých korpusů sem ještě patří dva korpusy věnované češtině v zahraničí. Jednak korpus BANÁT, kterému je věnována následující kapitola, jednak korpus texaské češtiny Texas Czech Dialect Archive vznikající v rámci projektu Texas Czech Legacy Project na The University of Texas at Austin.[19] Jedná se o korpus budovaný na základě již dříve pořízených nahrávek (z let 1970–2000). Jeho hlavním

[19] *About the Texas Czech Legacy Project* [online] [cit. 2014-08-07]. Dostupné z: <http://blogs.utexas.edu/txczech/texas-czech-dialect-archive/>.

cílem je nejen zachytit texaskou varietu češtiny a zpřístupnit ji dalšímu výzkumu, ale také zdokumentovat historii, příběhy a tradice této menšiny a v neposlední řadě vzbudit zájem o minoritní jazyky v zahraničí a získat tak podporu pro jejich zachování. Pro tvorbu korpusu jsou k dispozici stovky hodin nahrávek nepřipravených projevů, polořízených rozhovorů a vyprávění. Ambice autorů jsou poměrně velké – zpřístupnit nejen nahrávky s transkripcí, ale vzhledem k cizojazyčnému prostředí je doplnit i o volný překlad do angličtiny a morfologický popis.

Pražský závislostní korpus mluvené češtiny (PDTSC), který je součástí Prague Dependency Treebank of Spoken Language,[20] je syntakticko-sémanticky anotovaný korpus mluvené češtiny vybudovaný zejména pro účely pravděpodobnostního trénování a strojového učení s cílem plného porozumění mluvené řeči. Je založený na sběru z let 1993–1999 čítajícím 1260 hodin. Jednotlivé roviny analogické s Pražským závislostním korpusem, tj. za pomoci Funkčně generativního popisu, jsou analyzovány na rekonstrukci standardizovaného textu z mluvené řeči (Hajič et al. 2006).

4.4.3. Korpusy mediálních promluv a komunikace

Mluvenou češtinu v dnešních médiích zaznamenávají především dva korpusy vytvořené pracovníky Ústavu pro jazyk český AV ČR a ÚČJTK. Jedná se o korpusy DIALOG a Monolog, společný jim je kromě tématu stejný vyhledávací systém Dialogy.Org.

Korpus DIALOG[21] je multimediální korpus zahrnující nahrávky a přepisy diskusních pořadů českých televizních stanic,[22] který je nejen morfologicky anotovaný a lemmatizovaný, ale jeho součástí je i videozáznam, což z něj činí zástupce tzv. multimodálního korpusu, kde jsou dostupné nejenom přepisy a zvuk, ale i videozáznam, který přenáší vizuální informace. Ve verzi z roku 2012 (DIALOG 1.1) obsahoval 930 000 slov od 758 mluvčích ve 150 různých nahrávkách.

---

[20] *PDTSL – Prague Dependency Treebank of Spoken Language* [online] [cit. 2014-08-12]. Dostupné z: <http://ufal.mff.cuni.cz/pdtsl/>.

[21] *Korpus DIALOG* [online] [cit. 2014-08-12]. Dostupné z: <http://ujc.dialogy.cz/>.

[22] Nejzastoupenější (a s největší variabilitou) jsou pořady České televize, ale zastoupeny jsou i pořady vysílané stanicemi Nova, Prima a TV3.

Jazykový korpus Monolog[23] zahrnuje nahrávky a ortografické přepisy monologických veřejných projevů profesionálních mluvčích Českého rozhlasu. Poslední verze z roku 2013 obsahuje téměř 300 minut nahrávek a 40 000 slov. Korpus je lemmatizován i označkován.

4.4.4. Korpusy fonetické

Z fonetických korpusů je vhodné zmínit Pražský fonetický korpus (Prague Phonetic Corpus) vytvořený na Fonetickém ústavu FF UK, který se skládá z několika menších korpusů (pro detailní popis viz Skarnitzl 2010). Jednotlivé korpusy vznikly na základě nahrávek, které jsou segmentovány a přepsány stejným způsobem, a proto mohou položit základ fonetické analýze. Jedná se z velké části o čtené anebo polospontánní rozhovory pořízené v nahrávacím studiu, proto nahrávky dosahují více než dostatečné kvality bez dalších ruchů.

Na Fonetickém ústavu dále vznikl i korpus Minidialogy skládající se ze semispontánních krátkých dialogů (Volín et al. 2014) a právě je vytvářen korpus VASST (VAriabilita Skupin a STylů) (Weingartová et al. 2014), ve kterém jsou shromažďovány nahrávky z různých částí ČR zastupující různé mluvní styly (tj. spontánní dialogy, řízené dialogy, čtení textu, čtení vět atp.).

Na stejném pracovišti vznikl i korpus spontánní řeči The Nijmegen Corpus of Casual Czech,[24] který je spravován v nizozemském Nijmegenu (společně s podobnými korpusy francouzštiny a španělštiny). Skládá se z 30 hodin velmi kvalitních nahrávek rozhovorů 60 mluvčích rozdělených do trojčlenných skupin, které byly ortograficky přepsány (Ernestus et al. 2008).

[23] *Korpus Monolog* [online] [cit. 2014-08-12]. Dostupné z: <http://monolog.dialogy.org/>.
[24] *The Nijmegen Corpus of Casual Czech* [online] [cit. 2014-08-13]. Dostupné z: <http://www.mirjamernestus.nl/Ernestus/NCCCz/>.

5. SPECIALIZOVANÝ KORPUS BANÁTSKÉ ČEŠTINY

V této kapitole postupně popíši jednotlivé fáze, které stály za vznikem tohoto korpusu, případně které ještě v blízké budoucnosti proběhnou. Cílem není jen ukázat náročnost a specifika takového procesu, ale také ozřejmit některé volby, které v průběhu samotného vytváření ovlivnily finální podobu korpusu. Kapitola mapuje vývoj od roku 2010 až k době sepsání této práce. Nabízí také shrnutí úprav, které musí ještě proběhnout, než bude korpus BANÁT na podzim 2014 zveřejněn i pro širší vědeckou veřejnost mezi korpusy spravovanými ČNK.

Ač se může zdát, že se tato kapitola překrývá s jednou z kapitol v mé bakalářské práce (Vyskočilová 2012b: 22–27), ve skutečnosti je obsahově shodná pouze část úvodní (sběr materiálu), protože i dříve pořízené nahrávky byly opraveny podle nových pravidel, jak bude rozebráno níže.

V této části jsou rozlišeny tři různé verze korpusu BANÁT. První je pro účely této kapitoly nazvána BANÁT2011. Jedná se o korpus, který byl sestaven v roce 2012 z nahrávek získaných při prvním sběru dat. Sloužil zejména k předchozí analýze v mé bakalářské práci (Vyskočilová 2012b), byl lemmatizován a otagován. Skládá se z 21,5 hodiny nahrávek a přibližně 120 000 slov pouze banátských mluvčích.

Druhá verze je nazvaná BANÁT2014; zde jsou data z korpusu BANÁT2011 doplněna daty novými. Korpus není prozatím označkován a jedná se o pracovní a testovací verzi vytvořenou především za účelem analýzy v poslední části této diplomové práce. V korpusu BANÁT2014 zatím nebyly provedeny náročnější technické úpravy, pokud nebyly využitelné pro samotnou syntaktickou analýzu provedenou v následující kapitole. Příkladem může být např. to, že text není propojen se zvukem či že není možné vyhledávat v promluvách českých mluvčích.

Spojením zmíněných dvou korpusů vznikne finální korpus BANÁT, který bude zpřístupněn v nabídce korpusů ÚČNK. Jeho spuštění je plánováno na podzim roku 2014. Označení BANÁT používám také v případech, kdy mluvím o korpusu obecně, nikoliv o některé z jeho verzí. Označení BANÁT2011 a BANÁT2014 jsou používána pouze v případech, kdy je nutné striktně vymezit, o kterou z verzí korpusu se jedná.

Vzhledem k tomu, že by nebylo vůbec v silách jednotlivce shromáždit a zpracovat jazykový materiál ze všech šesti českých vesnic, materiál pochází jen z jedné vesnice, a to z Bígru. Pojmenování korpusu výrazem „BANÁT“ se může zdát zavádějící, ale důvodem byla nejen víra v jeho budoucí rozšíření, ale také snaha, pojmenovat korpus jasně. Název pro celou oblast je totiž poměrně známý a evokuje češtinu v Rumunsku, zatímco název „BÍGR“ by byl méně průhledný.

5.1. Východiska a cíle projektu

Myšlenka na vytvoření korpusu BANÁT se zrodila v roce 2010. Prvním impulzem byla nedostupnost materiálu pro výzkum banátské češtiny,[25] přestože se výzkumů uskutečnilo několik (viz kapitola 3) a během každého byl nasbírán různě rozsáhlý jazykový materiál.

Kromě potřeby získat materiál hrály roli i další faktory, jako snaha zachytit mizející variantu češtiny a zároveň ji zpřístupnit pro další výzkumy. V neposlední řadě také to, že cílem bylo umožnit porovnání banátské češtiny s jazykem na našem území, který je zachycený v korpusech ÚČNK řady ORAL.

Vzhledem k tomu, že srovnávání s korpusy ORAL bylo plánováno již od začátku celého projektu, měl být vytvořen korpus se srovnatelnými parametry. Nesporná výhoda byla v tom, že velké množství problémových otázek pro tvorbu korpusu (týkajících se zejména přepisu materiálů) bylo již vyřešeno. Navíc jsem se na tvorbě korpusů ORAL v průběhu několika let aktivně podílela (v oblasti nahrávání i přepisu) a mohla jsem proto při tvorbě korpusu BANÁT čerpat ze svých zkušeností.

Ve dvou zásadních věcech se korpusy BANÁT a ORAL liší. ORAL2008 je korpus sociolingvisticky vyvážený, korpus ORAL2013 je dokonce i reprezentativní.[26] Těchto parametrů nemůže korpus BANÁT dosáhnout a po vyhodnocení prvního sběru materiálu se o to nakonec ani nesnažil. Dotýká se to zejména počtu mluvčích, poměru mužů a žen, ale i počtu slov pronesených jednotlivými mluvčími. Vzhledem k tomu, že mezi obyvateli Bígru jsou zastoupeny především ženy důchodového věku, nebylo možné korpus vyvážit podle věku

[25] Výjimkou jsou nahrávky pořízené pro ČJA v 60. a 70. letech, které jsou dostupné v Dialektologickém oddělení Ústavu pro jazyk český AV ČR v Brně, přepis z nich pořídila Frnochová (2012). Vzhledem k tomu, že pro Šumici bylo k dispozici pouze 65 minut (Frnochová 2012: 27), nejedná se o celou sbírku pořízenou Skulinou. Dvě z těchto nahrávek jsou zároveň součástí CD *Jak se mluvilo v českých vesnicích v cizině* (Bachmannová a Jančák 2002).

[26] *Dostupné korpusy* [online] [cit. 2014-08-07]. Dostupné z: http://ucnk.ff.cuni.cz/struktura.php.

ani pohlaví, přestože to původně bylo záměrem. Mluvčí potřební k vyvážení korpusu ve vesnici často vůbec nejsou přítomni, anebo nejsou ochotní na projektu spolupracovat.

Druhou odlišností korpusu BANÁT2014 je fakt, že všechna data byla sesbírána na jednom místě. U korpusů řady ORAL takového počtu slov pro oblast pobytu v dětství mimo středočeskou oblast (tzn. zejména Praha dosahuje až novější, čtrnáctinásobně větší korpus ORAL2013.

První korpus byl sestavován se záměrem zkoumat mluvenou bígerskou syntax. Syntaktické jevy se až na výjimky špatně elicitují a navíc by nebylo možné srovnání s již dostupnými daty pro obecnou češtinu. Z toho důvodu bylo třeba nahrát maximálně spontánní rozhovory (stejně jako je tomu v korpusech ORAL).

V případě korpusu BANÁT bylo samotné nahrávání i jeho účel prozrazen před samotným nahráváním všem mluvčím. Motivací byla zejména snaha zachovat důvěru a otevřenost, se kterou mne krajané přijímali. Všichni mluvčí s nahráváním ústně[27] souhlasili a schválili také možnost dalšího vědeckého zpracování nahrávek. Z mých předchozích zkušeností s mluvenými korpusy vyplynulo, že i v takovém případě lze pořídit nahrávky, na nichž se mluvčí chová tak, jako by o nahrávání nevěděl. Každý mluvčí dříve či později na nahrávání zapomene a uvolní se, předešlou část nahrávky pak stačí oříznout a nezařazovat do přepisu.

Z lingvistického hlediska byl výběr vesnice Bígr motivován tím, že se výzkumu zdejší češtiny před mou bakalářskou prací věnovala podrobně pouze diplomová práce Vikové (1994). Druhou a neméně důležitou motivací bylo i to, že kvůli špatné dostupnosti vesnici navštěvuje mnohem méně turistů.

Již během své první cesty do Banátu v květnu 2007 jsem v Bígru navázala kontakty s místními, což mi při nahrávání v dalších letech výrazně pomohlo. V Bígru mi také byla nezištně zapůjčena chata RNDr. Ludmily Knappové, které bych tímto ráda ještě jednou poděkovala. Možnost pobývat právě v její chatě mi přinesla i další výhodu: jméno paní doktorky Knappové za mnou stálo při jednání s místními. Nebyla jsem tak během svých cest

[27] Ústně se získává i souhlas pro mluvené korpusy ÚČNK, z tohoto důvodu ho považuji za dostatečný.

považována za úplně cizího či neznámého člověka ani u lidí, se kterými jsem se setkala poprvé. Tento faktor byl obzvláště důležitý při prvním sběru, kdy se za mne postavil i místní učitel Mleziva. Krajané pak snáze a rychleji uvěřili faktu, že se jedná o nahrávání motivované jazykovědným zájmem, nikoliv snahou jakkoliv zesměšnit banátské Čechy či jejich jazyk.

5.2. Metoda sběru materiálu

Materiál sloužící jako podklad pro vytvoření korpusu BANÁT byl sesbírán během dvou výjezdů, které byly uskutečněny za podpory FF UK formou účelového stipendia. První z nich proběhl ve dnech 15. –25. dubna 2011, druhý výjezd se uskutečnil ve dnech 15.–24. března 2014. Doba pro oba výjezdy nebyla zvolena náhodně. Cílem bylo vybrat takové období, kdy budou vesnice dobře přístupné a zároveň v nich bude minimum turistů. Obojí se podařilo a během obou sběrů jsme byli přinejmenším v Bígru jediní Češi z České republiky ve vesnici. Na sběru materiálu se kromě mne podíleli i dva moji přátelé.[28]

Na tomto místě je vhodné podotknout, že pro získání různorodějšího materiálu by bylo vhodné jet do Banátu několikrát, ideálně několikrát ve stejném roce. Tímto způsobem měla možnost provést svůj výzkum pouze Viková, která se do Banátu v krátkém časovém horizontu dostala celkem čtrnáctkrát (Viková 1994: 1).

Tabulka 1 shrnuje počet hodin pořízeného materiálu vhodného ke korpusovému zpracování během obou výjezdů.[29]

	nahráno	**zpracováno**	**nezpracováno**
2011	22,5	22,5	0
2014	22	9,5	12,5
CELKEM	44,5	32	12,5

Tabulka 2: Přehled nasbíraného materiálu v Bígru (v hodinách)

V následující části se budu zabývat pouze problematikou nahrávání a zpracovávání dat nasbíraných v Bígru. Nahrávek pořízených v ostatních vesnicích není pro ucelenou analýzu dostatečné množství, ale domnívám se, že někdy v budoucnu mohou sloužit jako velmi malé referenční body či tvořit součást banátského korpusu v plném slova smyslu.

[28] Díky patří za doprovod při prvním sběru Adamovi Kulhavému a ve druhém Evě Volenové.
[29] Ve skutečnosti bylo nahráno přibližně o 30-40 % materiálu více, ten ale nebyl vhodný pro další zpracování.

Sběr v případě prvního výjezdu probíhal často na ulici, v případě výjezdu druhého povětšinou na návštěvě doma u mluvčích. Mluvčí tak nebyl vytržen ze svého prostředí a mohl se rychleji uvolnit a zapomenout na nahrávání. V nahrávkách se z toho důvodu někdy ozývají i rušivé zvuky v pozadí, jako např. zapnutá televize či zvuky domácích zvířat.

Přestože původním záměrem bylo nahrávat jen rozhovory banátských mluvčích bez přítomnosti mluvčích českých, již během prvního sběru se ukázalo, že to není možné (přinejmenším při takto krátkých výjezdech a pokud se jedná o nahrávání osobou mimo krajanskou komunitu). Místní samozřejmě nechtějí, aby se jich nahrávající pouze vyptával, ale naopak by rádi vedli přirozený rozhovor s někým, kdo je ve vesnici nový a s kým ještě nevyčerpali nejobvyklejší konverzační témata.

Přítomnost českého mluvčího v dialogu s jedním nebo dvěma banátskými mluvčími se může z mnoha důvodů jevit jako problematická, přestože se čeští mluvčí vědomě snaží nechávat banátským mluvčím co nejvíce prostoru. Především může dojít ke vzájemnému ovlivňování obou mluvčích, např. při výběru lexika, slovních obratů, témat hovoru apod. Z pozorování při přepisech a kontrolách nahrávek se ale zdá, že jsou to spíše čeští mluvčí, kteří podlehli některým vlivům banátské češtiny (intonace, banatismy, předložky s místními jmény apod.). U bígerských mluvčích k přejímkám docházelo zřídka, uchylovali se k nim jen ve chvílích, kdy mluvčí znal pouze rumunský či regionální výraz. V samotném korpusu budou oba typy promluv odděleny.

Při pořizování nahrávek jsem se snažila dosáhnout co nejkvalitnějšího záznamu, zejména s ohledem na skutečnost, že zvuková stopa bude přístupnou součástí zveřejněného korpusu a mohla by v budoucnu sloužit i pro dialektologickou či fonetickou analýzu.

Všechny nahrávky byly nahrány několika různými digitálními kompaktními rekordéry, vždy v nejvyšší možné kvalitě, tj. nekomprimovaný WAV se 16bitovou lineární pulzně kódovou modulací (LPCM) a vzorkovací frekvencí 44,1 kHz. V době psaní této práce vyšel článek Křivana a Zíkové (2014) věnovaný nahráváním v terénním lingvistickém výzkumu (zejména však pro fonetickou analýzu) a ráda bych proto na něj odkázala pro detailnější informace o technické stránce nahrávání, která je zde velmi dobře zpracovaná.

Během nahrávání byly průběžně zaznamenávány i nezbytné informace o mluvčích a nahrávkách, které jsou zčásti zachovány v atributech korpusu. U mluvčího se jedná o jméno (později kódované na identifikátor mluvčího), věk, místo původu a případně i vzdělání. U

každé nahrávky bylo zaznamenáno jméno nahrávajícího a dále datum a místo, kde byla nahrávka pořízena.

5.2.1. Mluvčí

Jak již bylo řečeno výše, u korpusu BANÁT se nepodařilo dosáhnout sociolingvistické vyváženosti ani podle věku, ani podle pohlaví. Všichni mluvčí za svůj mateřský jazyk považují češtinu, narodili se a vyrostli v Bígru, kde také strávili většinu svého života. Téměř všichni alespoň jednou navštívili Českou republiku a mají tam příbuzné. V aktuálně zpracovaných nahrávkách korpusu BANÁT2014 hovoří celkem osmnáct mluvčích, tři muži a patnáct žen.[30] Větší pestrosti se bohužel nepodařilo docílit.

Následující tabulka shrnuje počty mluvčích zahrnutých v korpusu BANÁT2014 a dělí je podle pohlaví a věku. V případě, že se stejný mluvčí zúčastnil nahrávání v obou letech sběru, je do tabulky zahrnut pouze jednou, primárně dle věku dosaženého při druhém sběru. Jeden mluvčí při prvním sběru patřil do jiné věkové kategorie (vyznačeno číslem v závorce), do celkového součtu je však započítán pouze jednou. V tabulce nejsou uvedeni tři čeští mluvčí, kteří pořizovali nahrávky a patřili po celou dobu sběru do kategorie 20–29 let. Vymezení kategorií bylo převzato z úzu korpusů řady ORAL ČNK.

	30–39 let	**40–49 let**	**50–59 let**	**60–69 let**	**70–79 let**	**80–89 let**	**CELKEM**
muži	0	0	1	1	1	0	3
ženy	1	1	3	4 (1)	5	1	15

Tabulka 3: Počet bígerských mluvčích rozdělených podle věku (BANÁT2014)

Objem nasbíraného materiálu a tudíž i počet slov vyslovených jednotlivými mluvčími ovlivnily ponejvíce dva významné faktory. První z nich byl omezený čas pro sběr a nemožnost kontrolovat objem nasbíraného materiálu jinak než poslechem či sčítáním délky nahrávek. Druhým byla rozdílná ochota či časová disponibilita jednotlivých mluvčích. Dochází kvůli tomu k nevyváženosti v počtu slov u jednotlivých mluvčích, respektive u mluvčích z jednotlivých věkových kategorií.

[30] Ve své bakalářské práci (Vyskočilová 2012b: 23) se zmiňuji o šestnácti ženách a čtyřech mužích, bohužel se ve dvou případech jednalo o duplicity, kdy byl jeden mluvčí označen omylem dvěma kódy.

Situaci v korpusu BANÁT2014 shrnuje následující tabulka. Předpokládám, že se rozložení finálního korpusu BANÁT může ještě změnit vzhledem k tomu, že z technických důvodů stále není zpracováno dalších téměř dvanáct hodin nahrávek.

	30–39 let	40–49 let	50–59 let	60–69 let	70–79 let	80–89 let	CELKEM
počet mluvčích	1	1	4	5 (1)	6	1	18
počet slov	3 875	1 932	49 237	25 888	115 328	14 406	210 747
slov na respondenta	3 875	1 932	12 309	5 177	19 221	14 406	11 708
poměr počtu slov	1,8 %	0,9 %	23,4 %	12,3 %	54,7 %	6,8 %	100 %

Tabulka 4: Poměr mluvčích a počtů slov v korpusu BANÁT2014

Tři čeští mluvčí, kteří nebyli zahrnuti do tabulky, pronesli v korpusu BANÁT2014 v součtu 58 606 slov, tj. přibližně čtyřikrát méně než mluvčí bígerští.

Dalšími hodnotami sledovanými ve všech korpusech ORAL jsou kromě věku, pohlaví, oblasti místa narození a pobytu také nejvyšší dosažené vzdělání a zaměstnání. Pokud je mi známo, většina respondentů je v důchodu (tento údaj byl doplňován bohužel zpětně).

Problematika vzdělání je výrazně složitější. Kdybychom se omezili na rozdělení používané ČNK, u všech mluvčích by byla uvedena hodnota „základní škola“. Nicméně jak již bylo zmíněno v oddílu 2.1.1, délka, obsah i jazyk vyučování v Rumunsku se v průběhu posledního století dost měnily. Informace o vzdělání jsem se pokusila zpětně doplnit při druhém sběru, většina respondentů uvedla, že odchodili čtyři třídy.

5.3. Metoda zpracování a přepisu nahrávek

Při prvním poslechu nahrávek byla posouzena jejich kvalita a zároveň byly nahrávky podle potřeby oříznuty v programu Audacity tak, aby pokud možno neobsahovaly ruchy vzniklé manipulací s diktafonem, dlouhá odmlčení či části projevu, v nichž se mluvčí soustředí na nahrávání a projev nepůsobí spontánně. V případě nových nahrávek byly delší nahrávky rozděleny, aby se s nimi lépe manipulovalo při vlastním přepisu. Údaj o tom, které sondy tvoří ve skutečnosti jednu dlouhou, zůstal zachován, aby bylo případně možné je znovu spojit.

Pro správu nahrávek a také pro formální kontrolu přepisů mi bylo umožněno používat mírně upravenou databázi Mluvka, která byla v ÚČNK využívána ke správě nahrávek pro korpusy řady ORAL. Databáze nejenže vyřešila formální kontrolu a zálohu dat, ale umožnila i jednoduché sdílení poměrně velkých nahrávek a odevzdávání přepisů bez komplikované emailové komunikace.

Změny v databázi provedené pro účely korpusu BANÁT jsou čistě technické. Především byl přidán mapovací soubor pro mluvčí (v původním BANÁTU2011 bylo použito kódové označení mluvčího, v přepisech pro ORAL se používalo postupné číselné označování mluvčích) a kontrolní skript, který kontroluje, zda přepsaná sonda odpovídá požadovaným formálním kritériím (oddělení pauz mezerou, výskyt velkých písmen v jiných než povolených případech apod.). Ke skriptu mi byl umožněn přístup a většinu pozdějších úprav jsem si byla schopná v programovacím jazyce Perl provést sama.

Zmíněná databáze Mluvka by nebyla potřeba, kdyby se na přepisu nahrávek (zejména v roce 2014) kromě mě nepodíleli i další spolupracovníci (z části financováni ÚČNK): Veronika Cikánová, Dominika Danielová, Michal Dudáš, Inka Dvořáková, Kateřina Gemrotová, Ivo Jelínek, Adéla Limburská, Anna Morávková, Pavla Morávková, Tomáš Lavička a Eva Volenová. Všem zmíněným patří velký dík, protože bez nich by nebylo jednoduché převést staré nahrávky do přepisu použitého pro účely této práce. Navíc přepis jednou osobou a kontrola druhou umožnily dvojí pozorný poslech a tím omezily počet chyb.

Pro přepis nahrávek byl použit stejný program jako v případě korpusů ORAL i BANÁT2011, a to volně dostupný program Transcriber. V dnešní době se pro přepisy častěji používá program ELAN. Přestože je převod přepisů z Transcriberu do ELANu technicky možný, nebyl proveden. Pro samotný přepis bígerských nahrávek s použitým nastavením by ELAN nepřinesl nic nového, je složitější na ovládání a navíc byly nahrávky korpusu ORAL2013 také přepsány v Transciberu (a již byly jednou v korpusu se zvukem zpracovávány).

Nahrávky byly přepsány podle mírně upravených „Pravidel přepisu sond pro korpus ORAL2013“,[31] která jsou v plném znění dostupná na webu ÚČNK. Jedná se tedy o transkripci „blízkou folkloristické (nikoliv tedy fonetické), tj. co nejblíže běžnému záznamu psanému, ale se speciálními úpravami pro účely počítačového zpracování podle úzu zavedeného v Českém národním korpusu“.[32] Poslední aktuální znění „Pravidel pro přepis sond pro korpus BANÁT“ je součástí této diplomové práce (viz příloha 1).

Kromě změny oproti korpusům ORAL spočívající v prozatímním vynechání anonymizace je další změnou pojetí segmentu, který výše zmíněná pravidla pro ORAL charakterizují jako „úsek přibližně odpovídající jednoduché větě, podřadnému souvětí nebo větnému fragmentu“; v mém pojetí byla však snaha o to, aby to byl „úsek tvořící přirozený syntagmatický, sémantický nebo prozodický celek (velmi přibližně odpovídající jednoduché větě či souvětí)“.

Oproti původnímu přepisu v korpusu BANÁT2011 je u korpusu BANÁT2014 nejvýraznější změnou přesné rozdělení do segmentů (velmi přibližně odpovídajících jednoduché větě) tak, aby bylo možné přepisy provázat se zvukem stejným způsobem, jako je tomu v korpusu ORAL2013; vyznačovány jsou i překryvy. Původní pokus o značení běžné interpunkce byl nahrazen pauzovou interpunkcí, která je s ohledem na častou neohraničenost mluvených projevů vhodnější.

V následujících tabulkách jsou shrnuty nejdůležitější body pravidel přepisu sond rozepsané v příloze 1. Číslo uvedené v prvním sloupci odkazuje ke konkrétní pasáži pravidel, v nichž jsou body podrobně popsány a případně doplněny příklady. Stav pravidel je platný pro přepis korpusu BANÁT2014, ale domnívám se, že již nedojde k žádné změně.

číslo pravidla	**shrnutí pravidel – zásady přepisu**
2.	Promluvy jsou rozděleny do segmentů (velmi přibližně odpovídajících jednoduché větě) tak, aby bylo možné přepisy provázat se zvukem stejným způsobem, jako je tomu v korpusu ORAL2013. Překryvy promluv dvou mluvčích jsou značeny.
3.	Je použita pauzová interpunkce, která je s ohledem na častou neohraničenost mluvených projevů vhodnější.
4.	Otázky jsou značeny otazníkem (*?*) a věty zvolací vykřičníkem (*!*).

[31] *Pravidla přepisu sond pro korpus ORAL2013* [online] [cit. 2014-08-08]. Dostupné z: https://ucnk.ff.cuni.cz/doc/Prepisovaci_pravidla_ORAL2013.pdf.

[32] *Pravidla přepisu sond pro korpus ORAL2013* [online] [cit. 2014-08-08], s. 1. Dostupné z: https://ucnk.ff.cuni.cz/doc/Prepisovaci_pravidla_ORAL2013.pdf.

6. a 7.	Označování neukončených promluv (....:) a přerušených výpovědí (...).
8.	Není značena přímá řeč.
10.	Vysvětlivky a poznámky k situaci jsou uvedeny v závorkách.
11.	Nerozluštěné úseky jsou značeny třemi pomlčkami (---).
12.	Prozatím nebylo přistoupeno k anonymizaci jmen.

Tabulka 5: Shrnutí obecných zásad přepisu sond

číslo pravidla	shrnutí pravidel
1.	Je dodržován standardní zápis, pokud je to možné, v případě více možných variant výslovnosti jsou zachycovány všechny.
2.	Pravidelné jevy, kde se i spisovná forma mluvená od formy psané odlišuje, se zapisují spisovnou formou zápisu.
3.	Zjednodušená výslovnost souhláskových skupin nebo jejich splývání (zejména na hranici slov) se nezachycuje.
4. a 5.	Ustálené i příznakové rysy běžné mluvy či rysy regionální se zachycují.
6.	Pokud může v jednom slově dojít k různé realizaci spodoby znělosti, je zachycována.
7.	Značíme kvantitu (fonologickou, emfatickou i emocionální), s výjimkou kvantity způsobené váháním mluvčího.
9.	Nedořečené slovo je značeno hvězdičkou za slovem (*).
10.	Interjekce zapisujeme rozděleně a podle toho, jak byly vysloveny.
11.	Responzní a hezitační zvuky jsou značeny následujícím způsobem: přitakání (*hmm*), zápor (*emem*), hezitační zvuk souhláskový (*mmm*) a samohláskový (*eee*).
12.	Zkratky jsou zapisovány podle způsobu výslovnosti.

Tabulka 6: Shrnutí vlastní transkripce

číslo pravidla	shrnutí pravidel - specifika
1.	Při označování mluvčích je používán speciální kód, který je unikátní pro každého mluvčího a obsahuje jeho číslo, pohlaví a věk.
2.	Jsou zavedeny speciální značky pro: • banatismy (# před slovem) – slova cizího, regionálního původu, popřípadě i rumunská slova; přepisovány jsou podle toho, jak je slyšíme • banátské „hovorové tvary“ (*&* před slovem) – tvary běžně používané v Banátu, které nepatří do běžné obecněčeské zásoby, slova, která mají jinou koncovku či nestandardní výslovnost • netypické obecněčeské hovorové tvary (@ před slovem).
3.	Pro slova, u kterých si přepisující nejsou zcela jisti, že rozuměli, se používá značka *(?)*.
4.	Anonymizační značky (pro české mluvčí): *XA(počet vteřin)* – anonymizace z důvodu citlivého obsahu hovoru; *XM(počet vteřin)* – monolog českých mluvčích, banátský mluvčí přítomen, ale následuje v dalším segmentu např. dlouhé odmlčení; *XN(počet vteřin)* – hovor českých mluvčích v nepřítomnosti banátského mluvčího.

Tabulka 7: Shrnutí specifik korpusu BANÁT

Při anotování speciálních značek byla výhoda v tom, že přepisující, kteří nejprve segmentovali staré nahrávky, už viděli značky # a *&* v praxi a mohli proto označovat podobné případy. Byli instruováni, aby raději značku nepoužívali, pokud si jejím užitím nejsou jisti. Značky byly použity s ohledem na budoucí možné značkování, aby jimi označená slova nesnižovala úspěšnost značkovacích nástrojů.

Značky jsem doplňovala a případně opravovala při kontrolách, které jsem prováděla sama, čímž byla zajištěna konzistence. Ukázalo se, že značit „netypické obecněčeské tvary“ nemá smysl, protože hranice byla příliš vágní a nebylo v mých silách udržet konzistentní značení. Zároveň jsem během kontroly doplňovala anonymizační značky pro české mluvčí,

které se do pravidel přidaly později. Užila jsem je z toho důvodu, že v případě značky pro vynechání (*(...)*), která se používala v korpusu BANÁT2011, nebylo zřejmé, co mělo být důvodem nepřepsání takové promluvy ani jaká je její skutečná délka.

Vzhledem k tomu, že jsem si ponechala původní přepis odevzdaný přepisujícími, kteří nikdy s banátskou češtinou nepřišli do styku a nahrávání se neúčastnili (až na jednu výjimku) a ostatně ani s přepisy samotnými neměli zkušenosti (až na jednu výjimku), je při srovnání finálních přepisů možné sledovat jevy, které přepisujícím činily potíže. Tyto informace mohou po podrobnější analýze sloužit jako podklad pro studii zabývající se přepisy mluvených projevů provedenými takovým mluvčím (nebyl přítomen a nemá zkušenost s přepisováním) ve srovnání s mluvčím, který nahrávání přítomný byl, ale s přepisy přesto zkušenosti nemá.

Klasickými chybami při přepisování mluvených materiálu jsou chyby z nepozornosti jako překlepy, špatné odposlechnutí výslovnostní varianty (např. slovo *dělala* může být vysloveno i redukovanou formou *ďála/děala* apod.), nenaznačení pauz či špatné dělení segmentů. Chybami specifičtějšími pro tento korpus bylo často to, že přepisující slyšeli pojmenování či výrazy, které jim byly neznámé, a v případě nějakého ruchu v pozadí se snažili najít v promluvě známá slova (např. místo *makar ukradnutý* přepsáno *má káru* ---). Tomuto problému se dá v některých případech předejít díky tomu, že většina rozhovorů primárně má svůj význam, a je proto lepší takové slovo nepsat anebo alespoň zaznamenat s poznámkou (např. *kovářovic kobyla chodí bosa* zapsaná jako *kovářovic kobyla chodí psa*). Druhým poměrně zajímavým faktem bylo, že se přepisujícím často zdála podezřelá (vyznačeno *(?)*) slova, která mají v dnešní češtině velmi expresivní význam, ale v banátské češtině jsou to slova poměrně běžná. Pokud se v přepisu takové slovo objevilo v běžné neexpresivní promluvě, přepisující váhal, jestli slyší dobře, protože se mu to v projevu staré paní zdálo nepatřičné (např. *voni mají rodinu každej mají haranta*).

5.4. Vytvoření korpusu

Po kontrole nahrávek je možné přistoupit k tvorbě samotného korpusu. Jednotlivé přepisy nahrávek jsou standardně uložené ve formátu TRS, který je sice asociovaný s Transcriberem, ale ve skutečnosti se jedná o formu XML. Z těchto souborů bylo potřeba vytvořit vertikálu, která je nutná pro publikaci korpusu pro korpusová rozhraní ÚČNK.

Zjednodušeně řečeno, vertikála odpovídá prostému textu ve formátu jedno slovo na řádek (a jako takováto jednotka vystupuje řádek coby jeden token při vyhledávání

v korpusu), který je doplněn několika speciálními značkami značícími začátek dokumentu či změnu mluvčího. Vzhledem k tomu, že v korpusu nebyl nalezen vhodný skript pro konverzi přepisů, bylo třeba ho napsat znovu. V souvislosti s tím bylo nutné vyřešit několik technických problémů, jako je převod překryvů či způsob zpracování poznámek a projevů českých mluvčích. Vše bylo vyřešeno stejně jako v korpusu BANÁT2011 či ve starších korpusech řady ORAL, tj. překryvy byly rozděleny na jednotlivé mluvčí, poznámky v kulatých závorkách tvořily jeden řádek/token, stejně tak promluvy českých mluvčích, které byly opatřeny hranatými závorkami, a opět jeden segment odpovídá řádku/tokenu. Odstraněny byly prozatím i přidané znaky #, *&* a @. Samotný skript v Perlu je i s ukázkou vertikály součástí přílohy 2.

V případě korpusu BANÁT2011 byl přepis dále opatřen lemmaty a morfologickou anotací (viz. Vyskočilová 2012b: 24–25). Korpus BANÁT2014 přes veškerou snahu oanotován nebyl, zejména kvůli technickým problémům.[33]

Ani korpus ORAL2013, se kterým jsou v této diplomové práci bígerské jevy porovnávány, anotovaný zatím není, takže stejně bude třeba vyhledávat jednotlivé tvary pomocí seznamů. K dispozici není ani pracovní verze takovéto anotace, i když se v budoucnu se značkováním mluvených korpusů počítá. Navíc pokud by mělo jít o anotaci přesnou, bylo by s ohledem na velké množství odlišných bígerských slov potřeba zasahovat ručně. Jenom součet slov opatřených značkou & anebo # činí téměř 1 700 typů, přičemž tvarů hovorových, se kterými současné značkovací nástroje nepočítají, je jistě mnohem více. Na ruční zásah (který by byl velmi časově náročný) nezbyl čas, protože i samotné zpracování přepisů se zpozdilo – nebylo vůbec snadné najít vhodné přepisující. Ačkoli se tedy s přepisy mělo začít již koncem roku 2013, první z nich byly odevzdány až v březnu 2014.

Nakonec byl vytvořen korpus BANÁT s parametry uvedenými v následující tabulce:

	BANÁT2014	
	bígerští mluvčí	**čeští mluvčí**
počet pozic (tokenů)	265 703	
počet pozic (tokenů) bez interpunkce a dalších značek	210 747	(58 606)[34]

[33] Tímto bych ještě jednou ráda poděkovala kolegům RNDr. Haně Skoumalové, Ph.D., a Mgr. Pavlu Vondřičkovi, Ph.D., za čas, který strávili nad nakonec neúspěšnými pokusy.

počet slovních tvarů (wordů)	16 587	-
počet promluv	29 632	17 005
počet unikátních (různých) mluvčích	18	3
délka nahrávek	32 hodin	

Tabulka 8: Parametry korpusu BANÁT2014[35]

5.5. Zveřejnění korpusu BANÁT

Korpus vytvořený metodou popsanou v předchozí podkapitole je zatím dostupný pouze na vyžádání. Podmínka pro částečné financování tvorby korpusu BANÁT ze strany ÚČNK byla ta, že se korpus zveřejní ještě letos mezi specializovanými mluvenými korpusy. Prozatím se počítá s jeho zveřejněním na podzim 2014.

V publikovaném korpusu budou všechny bígerské nahrávky uvedené v tabulce 2. Přepis bude propojen se zvukovou stopou (stejně jako je tomu u korpusu ORAL2013). Na obou webových rozhraních NoSketch Engine a KonText bude možné ke každému zobrazenému segmentu přehrát danou část zvukové stopy. Dále bude umožněno prohledávat promluvy jak bígerských, tak českých mluvčích.

Pokud se to ukáže jako nutné, proběhne ještě třetí kontrola nahrávek. K vyřešení zbývá jen několik detailů, jako je otázka opětovného spojení dat, využití použitých speciálních značek (možnosti jejich zachování někde v korpusu jako další atribut), přidání jména vesnice jako charakteristiky mluvčího společně s měsícem a rokem sběru materiálu (prozatím je zachováno v názvu sondy). Nejpalčivější je otázka anonymizace mluvčích zmiňovaných v rozhovorech – na jednu stranu by to snad bylo vhodné, na druhou stranu stojí za zvážení, zda je identifikace skutečně možná, když mluvčí ve většině případů zmiňují křestní jméno spolu s příjmením, které je však stejné pro čtvrtinu vesnice.

5.6. Možné rozšíření korpusu BANÁT

Korpus BANÁT svým názvem přímo vybízí ke svému rozšíření o další nahrávky. Takových nahrávek bylo za posledních pár let pořízeno několik (kromě těch, které jsem nahrála sama). Nahrávky mi po roce 2012 nabídlo několik lidí. Digitální jsou (a povětšinou i s přepisem) nahrávky Adély Frnochové a Karoliny Haiderové, dále mi byly nabídnuty

[34] Do celkového počtu pozic jsou promluvy českých mluvčích započítány ne podle tokenů, ale podle počtu promluv. Kvůli zamezení vyhledávání v českých promluvách byla celá promluva označena jako jeden token.
[35] Struktura tabulky je vychází z přehledových tabulek používaných ÚČNK.

nahrávky pořízené pro projekt Post Bellum.[36] Vilma Anýžová (roz. Viková) nabízela velké množství MC kazet s nahrávkami z 90. let. Hlavním problémem však je, že přepisování či pouhá úprava již přepsaných nahrávek zabere velké množství času a bez dalšího financování je takový projekt nemožný. Druhou možností, jak korpus rozšířit, by bylo případné další zájemce o psaní podobné graduační práce na téma jevů banátské češtiny motivovat k tomu, aby k přepisu materiálu přistupovali výše popsaným způsobem, přepisy pak zařazovat do korpusu a výměnou nabídnout vyhledávací rozhraní.

[36] *Post Bellum* [online] [cit. 2014-08-12]. Dostupné z: <http://www.postbellum.cz/>.

6. LINGVISTICKÁ ANALÝZA

V této kapitole se zaměřím na lingvistickou analýzu přivlastňovacích zájmen, záporných zájmen a příslovcí a příklonek v bígerské češtině na pozadí korpusu mluvené češtiny ORAL2013. Jejich výběr byl podložen výsledky předchozího výzkumu (Vyskočilová 2012b: 27–61, 2013), ve kterém se právě tato slova ukázala být charakteristické pro bígerskou češtinu. Zdrojovými daty pro výzkum jsou korpus BANÁT2014 (pro podrobnější popis viz kapitola 5), který reprezentuje stav jazyka v Bígru, a korpus ORAL2013 pro obecnou češtinu.

Pro účely srovnání vybraných jevů bígerské syntaxe s obecnou češtinou na našem území byl z korpusu ORAL2013 vytvořen subkorpus z oblastí předpokládaného původu bígerských krajanů, tj. z jihozápadočeské a středočeské oblasti. Takto vytvořený korpus obsahuje 900 170 textových slov (pro srovnání: velikost korpusu BANÁT je 210 747 textových slov). Pro přehlednost bude v následujících podkapitolách tento subkorpus označován jednoduše „ORAL“, případně bude odkazováno na stav „obecné češtiny“.

Všechny níže zmiňované jevy jsou v korpusech vyhledávány podle slovních tvarů (*word*), protože jiný způsob vyhledávání není v neoznačkovaných korpusech možný. Oproti lemmatu však může nastat problém, že jsou vybrané tvary homonymní s nějakým jiným slovem (viz níže příklad slova *má*), proto je třeba třídit výsledky obzvláště pečlivě a počítat dále jen s tvary, které byly opravdu hledány. Analýza dat proto kvůli kontrole probíhala dvakrát s časovým odstupem a jevy byly pokaždé seřazeny v jiném pořadí.

Nalezené hodnoty pro hledané jevy a jevy s nimi související jsou dále statisticky vyhodnocovány. V případě, že byl počet výskytů jevu příliš velký a nebylo ho kvůli časové náročnosti možné protřídit celý ručně, byl pro úplnost dopočítán výskyt hledaného jevu v populaci. K tomuto výpočtu byl použit interval spolehlivosti pro jednorozměrnou veličinu – exaktní metoda (*Confidence interval for single proportion – exact method*), který je založen na binomickém rozdělení, kterému níže zpracovávaný typ dat odpovídá. Interval spolehlivosti určí, v jakém rozmezí se mohou naměřené hodnoty pohybovat se zvolenou pravděpodobností (zde 95 %) (Milička, v tisku; Wallis 2013).[37]

[37] K výpočtům byl použit program aStat [online] [cit. 2014-08-13]. Dostupné z: <https://play.google.com/store/apps/details?id=org.twbbs.astat>.

Pro porovnání vzájemného vztahu sledovaného jevu ve vybraných jazykových varietách (v bígerské a obecné češtině) byl použit Fisherův test (*Fisher's exact test*) (Fisher 1922).

6.1. Zájmena přivlastňovací

V této podkapitole bych ráda podrobněji rozebrala používání zvratného přivlastňovacího zájmena *svůj* v mluvených spontánních projevech. Ve své bakalářské práci (Vyskočilová 2012b: 57–58) v návaznosti na předchozí výzkumy (Ciplea 1971: 218; Frnochová 2012: 78; Utěšený 1964b: 30) potvrzuji, že v bígerské češtině dochází k záměnám zvratného zájmena *svůj* a tvarů zájmena *můj* a *jeho*. Ciplea (1971: 218) tuto záměnu považuje za vliv rumunštiny; níže uvádím jeho příklady (1)–(3), které doplnil i paralelní rumunskou větou/frází (v závorce).

(1) *celej muj život sem ďelal* (*toată viaţa mea am lucrat*)[38]
(2) *poslal jeho sekretáře* (*a trimis pe secretarul lui* (2. pád sg.))
(3) *von tadi bil celej jeho vjek* (*viaţa lui*)

K potvrzení této hypotézy docházím v předchozím výzkumu docházím srovnáním užití lemmat *můj*/*jeho* na místě *svůj* v korpusu BANÁT. Na základě výsledků dopočítaných Studentovým t-testem[39] získávám frekvenci jevu v korpusu 50±11/518.

Problematika přivlastňovacích zájmen v bígerské češtině je ale mnohem složitější. Zmíněný výzkum měl několik nedostatků; jednak jsem řešila obě zájmena dohromady (*můj* i *jeho*), jednak záměna nebyla porovnávána se situací v obecné češtině. Vzhledem ke komplexitě celého jevu jsem nyní omezila rozsah sledování opět pouze na záměny zájmena *svůj* za tvary *můj* a *jeho/její/jejich*.

Teoretické pozadí této podkapitoly je přejaté z novějšího článku Čmejrkové (2003), který je věnován zvratnému zájmenu *svůj* a jeho záměnám s osobními zájmeny (v češtině na našem území). Článek je částečně postaven na analýze předchozích studií a následně je na rozdíl od dalších podobných studií doplněn korpusovými příklady z mluveného i psaného korpusu (přestože obvykle není uvedeno, ze kterého z nich příklady pochází). Čmejrková nad příklady navíc souhlasí s tím, že z různých (dále rozebraných) důvodů může být reálně užití

[38] Všechny příklady přejaté z dřívějších výzkumů jsou uváděny v původní transkripci.
[39] Vycházím z pěti náhodných vzorků po padesáti výskytech.

jiné, než by se dle spisovných pravidel očekávalo. Právě jejím rozdělením se budu řídit při analýze korpusových dat a porovnávání bígerské a české situace.

Obecně platí, že pokud je posesor v podmětu, lze přivlastňovací zájmeno *můj/tvůj/jeho* (v libovolném rodě) nahradit zvratným zájmenem *svůj/svá/své*. Do užívání těchto zájmen se navíc promítá postoj mluvčího, a to tak, že zájmeno *svůj* je sémanticky vyprázdněné a mluvčí proto dává přednost osobnímu přivlastňovacímu zájmenu.

6.1.1. *Svůj* a *můj*

Ohledně záměny *svůj* za *můj* Čmejrková (2003: 190) zobecňuje, že „[v]e spontánních mluvených projevech se setkáváme s vysokou frekvencí přivlastňovacího zájmena *můj* v pozicích, kde bychom podle pravidla o reflexivizaci při přivlastňování gramatickému podmětu očekávali zvratné zájmeno *svůj*“. Tuto hypotézu se pokusím dále potvrdit nebo vyvrátit na obou korpusech.

Podle jejího shrnutí se jedná zejména o kontexty, ve kterých se mluvčí zmiňuje o sobě, případně o věcech, které s ním úzce souvisí. Příklady Čmejrková (2003: 190–192) dělí následujícím způsobem: (4) integrita své vlastní osobnosti a existenčního pole; (5) přehlédnutí dosavadního života; (6) rodina; (7) peníze; (8) přátelé, kolegové a hosti; (9) vlast; (10) fyzické výkony a materiální výtvory; (11) výtvory duchovní.

(4) *To jsem na mou duši zapomněla.*
(5) *Podrobil jsem se té nejtěžší zkoušce mého života.*
(6) *To jsem dostal od mé ženy.*
(7) *Budu živit rodinu z mýho platu.*
(8) *Řekl jsem to mému kamarádovi.*
(9) *K mé vlasti mám vztah nejlepší.*
(10) *Budu zkoušet vylepšit můj skok.*
(11) *Když přednáším ten můj předmět.*

Jako motivaci pro užívání zájmena *můj* místo *svůj* Čmejrková (2003: 192–193) uvádí nejen zmiňovanou expresivitu, ale to, že se mluvčí ocitne v situaci, kdy při změně podmětu zapomene změnit zájmeno či má strach z nesprávného užití. V mluvené řeči je *můj* v tomto případě používané nejspíš zejména proto, že je subjektivnější a důraznější než *svůj*.

Tvary zájmena *můj* byly v korpusech ORAL i BANÁT2014 hledány pomocí výčtu, který je čerpán ze seznamu uváděného Cvrčkem (2010: 214) pro spisovné i hovorové tvary zájmen. Pro hledání jsem vytvořila paradigma skrývající se za lemmatem *můj*, tj. kompletní seznam sjednocených rodů a obou čísel bez duplicit: *má*, *mé*, *mého*, *mejch*, *mejma*, *mém*,

mému, *mí*, *moje*, *mojeho*, *mojemu*, *moji*, *mojí*, *mojim*, *mou*, *muj*, *můj*, *mý*, *mýho*, *mých*, *mym*, *mým*, *mýma*, *mými, mýmu*. Přehled výsledků korpusového dotazu je shrnut v tabulce 9.

Níže následuje výčet všech jevů nalezených mezi náhodnými vzorky z korpusu BANÁT2014 (12)–(19) a jeden příklad z korpusu ORAL (20), žádný další výskyt v korpusu ORAL překvapivě nalezen nebyl. Hledané zájmeno je v příkladech vyznačeno tučně, nutný kontext vyplývající z předchozích promluv je doplněn v hranatých závorkách. Příklady lze zařadit do kategorií, ve kterých podle zmiňované Čmejrkové *svůj* není užíváno; tematicky se týkají stavu osobnosti (12), rodiny (13) a (14), majetku (14)–(16)(17), peněz (17)–(18) a fyzických výkonů (19).

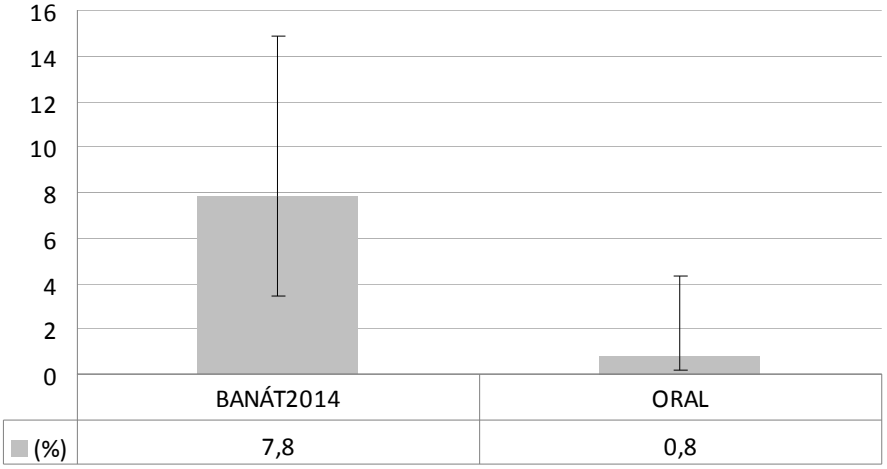
(12) *poďte já vám vokážu nahoře . jaký to mám tu **mojí** chudobu ..*
(13) *přínesu od mo* od **mojí** .. eee holky tam z Dolácka*
(14) *já je nekupuju . já mám **moje** [vejce]*
(15) *až to přivezou ičko ... jo .. budu mít **moje** [kuřata] . ...*
(16) *to muším dát **mojim** pipinkám .. nate*
(17) *já tady budu mít peníze **moje** a nemám prej se žádným nic*
(18) *sem řikal každej měsíc můžu koupit jednu koterici .. na **moje** peníze sem si řekl*
(19) *co sem se já nadojila z **mejma** rukama mmm mlíka že . by moh tect jeden potok*
(20) *že já teda sem takovejdle kontakt neměla ani s **mym** vlasnim mužem*

		BANÁT2014	**ORAL**
počet výskytů hledaného lemmatu		958 (ipm[40] 3 605,53)	2 754 (ipm 2 670,68)
velikost náhodného vzorku		250	750
z toho	**tvar slovesa *mít***	146	622
	můj	**94**	**125**
	možné *svůj*	**8 (7,8 %)**	**1 (0,8 %)**
	ostatní	2	2
konfidenční interval pro možné *svůj* v populaci *můj*		3,447–14,87 %	0,2009–4,343 %

Tabulka 9: Výsledky hledání lemma *můj*

Tabulka 9 shrnuje informace pro korpusy BANÁT2014 a ORAL, dále obsahuje informaci o velikosti vybraného náhodného vzorku. Jeho analýzou byly dané tvary rozděleny do čtyř kategorií: (a) případ, kdy se jedná o tvar slovesa *mít* v 3. osobě singuláru, (b) běžné tvary zájmena *můj*, (c) tvary zájmena *můj* ve kterých by měl/mohl být použit tvar zvratného *svůj*; (d) kategorie „ostatní“, která obsahuje tvary zájmena, kolem kterých byl nedostačený kontext (tj. nerozluštěné úseky) anebo se jednalo o jiný slovní druh (zde citoslovce *mé*). Výsledek společně s vyznačeným konfidenčním intervalem je znázorněn v grafu 1.

[40] Zkratka ipm (*instances per million*) zachycuje relativní frekvenci jevu (tj. výsledku dotazu), vyjadřuje pravděpodobný počet výskytů tohoto jevu na milion (vztaženo k celému korpusu).



Graf 1: Potenciální užití *svůj* místo *můj*

Spočítáním Fisherova testu je získána *p-value* = 0,005541 < 0,05, která dokazuje, že oba zkoumané vzorky dat jsou na sobě nezávislé, což v tomto případě znamená, že situace v bígerské češtině se odlišuje od stavu v češtině obecné. S ohledem na to, že předchozí zmínky o jevu pocházely z jiných banátských vesnic, je možno předpokládat, že situace bude obdobná i v banátské češtině jako celku.

6.1.2. *Svůj* a *jeho*, *její*, *jejich*

Podle Čmejrkové (2003: 194–196) je konkurence *svůj* a *jeho*, *její*, *jejich* oproti přivlastňování 1. a 2. osobě specifická v tom, že funguje jako anafora a má proto větší dosah na srozumitelnost výpovědi. Z kontextu bývá často význam jasný v případě mluvčího a adresáta (21), v případě reference o třetí osobě může nastat problém s určováním antecedentů – povrchový a méně pravděpodobný antecedent nebo antecedent skrytý v hloubkové struktuře (22). Klíčem k rozluštění může být kontext, viz (23) – v případě situace „na letišti" je význam jasný. V případě chybějícího kontextu může naopak neužití zájmena *svůj* vést k pochybnostem o referenci (24). Důvodem neužití reflexivního zájmena může být odstup mluvčího od sdělení (25), stejně jako u výše zmíněného *můj*.

(21) *Slyším tě zpívati svou oblíbenou píseň.*
(22) *Slyšel ho zpívat svou oblíbenou píseň.*
(23) *Brzo jsem viděla člověka s cedulkou se svým jménem.*
(24) *Šel do kina s jeho ženou.*
(25) *Domorodci Vás ničím nevyvedli z míry, jejich způsobem života, jejich komunikací s Vámi?*

Tvary zájmena *jeho*, *její*, *jejich* jsem hledala stejným způsobem jako výše uvedené *svůj*. Seznam byl čerpán z přehledu ve Cvrčkově mluvnici (2010: 216) pro spisovné tvary. Hledáno

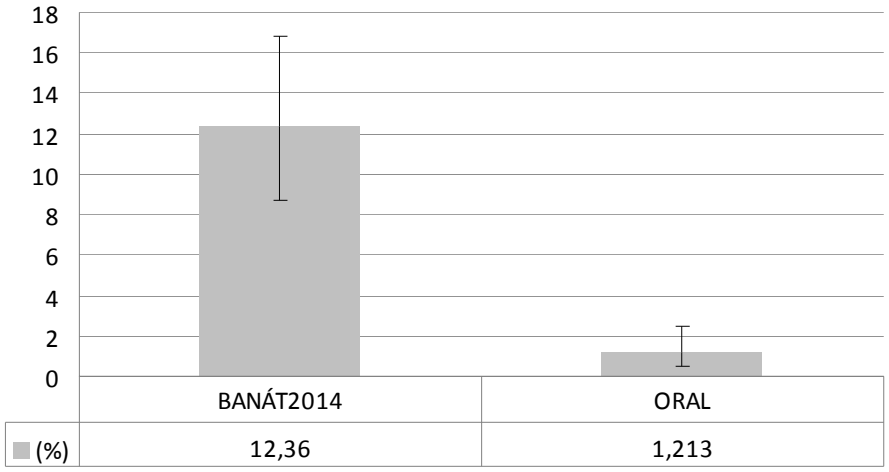
bylo za pomoci těchto tvarů: *jeho*, *její*, *jejího*, *jejich*, *jejích*, *jejim*, *jejím*, *jejíma*, *jejími*, *jejímu*. Výsledky jsou uspořádány v tabulce 10. Níže uvádím několik příkladů, kompletní seznam viz příloha 3. Z uvedených příkladů vyplývá, že motivace pro užití *jeho*, *její*, *jejich* není odlišná od kategorií uvedených v předchozím oddíle (6.1.1). Ve většině případů se jedná o rodinu.

(26) *to je Katka s **jejím** manželem*
(27) *s námi mluvila tak tady ho [stavení] má . jako po **jejích** rodičích*
(28) *dyž tam přišel s **jeho** tátou a ..*
(29) *na **její** nemoc . dostala důchod víš drobet*
(30) *byla já nevim v kerý zemi byla s ňákým druhým **jejím** milencem*
(31) *Mi[l]ada prý chodí s **její** mámou někam po špitále zase jo*
(32) *no děti chodí . ty chodí s těma **jejich** řechtačkama ..*
(33) *jo Rumun asi ne Rumuni ty maj tu **jejich** víru ..*

	BANÁT2014	**ORAL**
počet výskytů (populace)	275 (ipm 1 034,99)	577 (ipm 559,54)
jeho, její, jejich	241	570
možné *svůj*	34 (12,36 %)	7 (1,213 %)
konfidenční interval	8,716–16,85 %	0,4891–2,484 %

Tabulka 10: Výsledky hledání lemma *jeho*, *její*, *jejich*

I v tomto případě spočítání Fisherova testu a získání *p-value* = 3,052e-11 < 0,05 vyplývá, že situace v bígerské češtině se odlišuje od stavu v češtině obecné. Na následujícím grafu je vidět, že rozdíl je signifikantní nejen statisticky. Porovnáním dolního limitu konfidenčního intervalu bígerské češtiny a horního intervalu ORALU je vidět, že výskyt tohoto jevu v korpusu BANÁT2014 je nejméně třikrát vyšší než v korpusu ORAL.



Graf 2: Potenciální užití *svůj* místo *jeho*, *její*, *jejich*

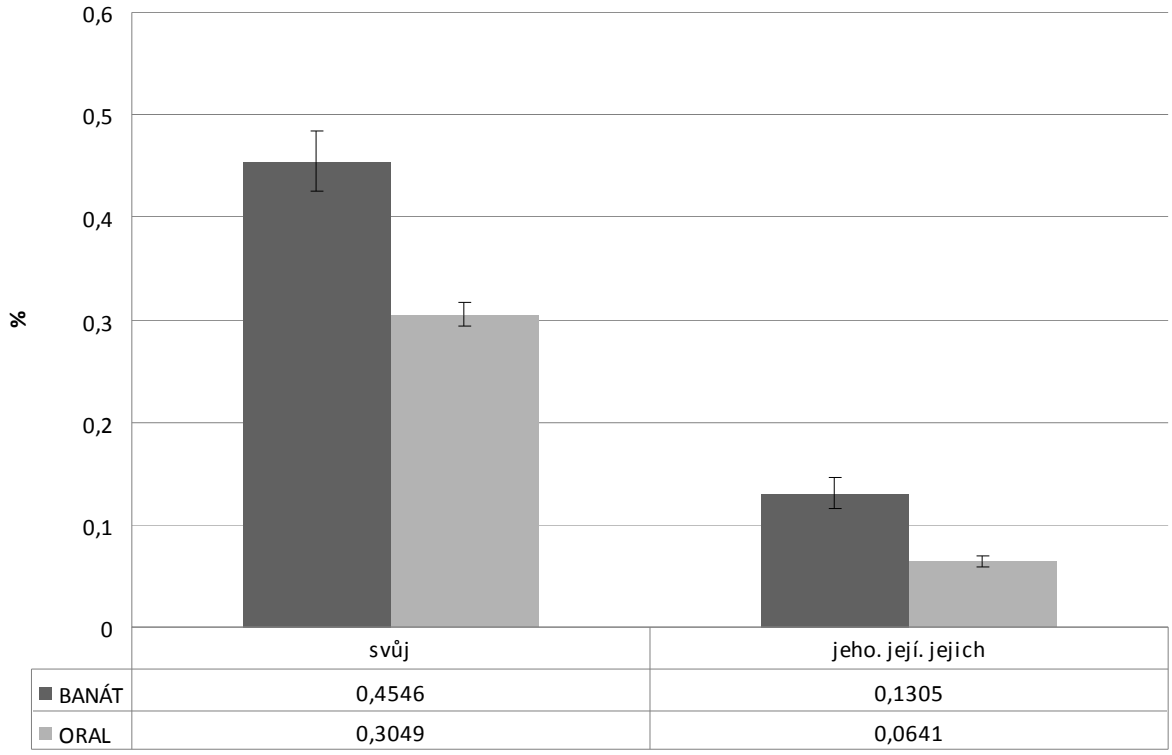
6.1.3. Srovnání frekvence zájmen *můj*, *jeho*, *její* a *jejich*

Z předcházejících dvou výzkumů byla také získána data o celkovém počtu zájmen *můj, jeho*, *její* a *jejich* v obou varietách češtiny. Data byla sloučena do následující tabulky:

	BANÁT2014	**konfidenční interval**	**ORAL**	**konfidenční interval**
počet slov v korpusu[41]	210 747	-	900 170	
svůj	958 (0,4546 %)	0,4263–0,4842 %	2 745 (0,3049 %)	0,2937–0,3165
jeho, její, jejich	275 (0,1305 %)	0,1155–0,1468 %	577 (0,06410 %)	0,05898–0,06955 %

Tabulka 11: Srovnání frekvence zájmen *svůj*, *jeho*, *její*, *jejich*

Fisherův test v obou případech poskytne *p-value* = < 2.2e-16, z čehož plyne signifikantní rozdíl mezi bígerskou a obecnou češtinu, který je vidět i na grafu 3. Četnější užívání zájmen *svůj*, *jeho, její, jejich* může souviset s tématy hovorů obou korpusů, ale neméně zajímavé by v případě bígerské češtiny bylo prozkoumat, zda by se mohlo jednat o vyjádření určitosti (stejně jako v rumunštině), tak jako v případě častějšího užití *jeden*, *jedna*, kdy se jedná o kalk rumunského *un*, *una* jako neurčitého členu (Vyskočilová 2012b: 56–57).



Graf 3: Srovnání frekvence zájmen *svůj*, *jeho*, *její*, *jejich*

[41] V případě korpusu BANÁT2014 se jedná pouze o počet slov bígerských mluvčích.

6.2. Záporná zájmena a zájmenná příslovce

V této podkapitole se zaměřím na záporná zájmena a zájmenné příslovce. V prvním oddíle jsou analyzovány výrazy *nic* a *nikerak*, které slouží v banátské češtině k vyjadřování zesíleného záporu. Druhý oddíl tvoří analýza šest zájmen a zájmenných příslovcí společných oběma korpusům, a to *nikam*, *nikde*, *nikdo*, *nikdy*, *nikoho* a *nikomu.*

6.2.1. Zesílený zápor – *nic* a *nikerak*

Ve své bakalářské práci (Vyskočilová 2012b: 48) potvrzuji v souladu s dalšími studiemi (Frnochová 2012: 73; Haiderová 2007: 90; Utěšený 1962: 205, 1964b: 30; Viková 1994: 32) užívání dvou totálních kvantifikátorů za účelem vyjádření zesíleného záporu (ve smyslu *vůbec*), a to *nic* a *nikerak*.

V ORALu nebylo po vyhledání korpusového dotazu „nik…*“ nalezeno žádné jiné zájmeno nebo příslovce než šest rozebíraných v následujícím oddíle, v korpusu BANÁT2014 bylo nalezeno několik různých variant od většiny z nich. Z toho vyplývá, že *nikerak*, kterého bylo nalezeno celkem 42 výskytů (navíc po jednom případu varianty *nikerakž* a *nikeraš*) je rysem čistě typickým pro bígerskou češtinu. Toto zájmenné příslovce vyjadřuje zesílený zápor ve všech případech, uvedu alespoň několik z nich:

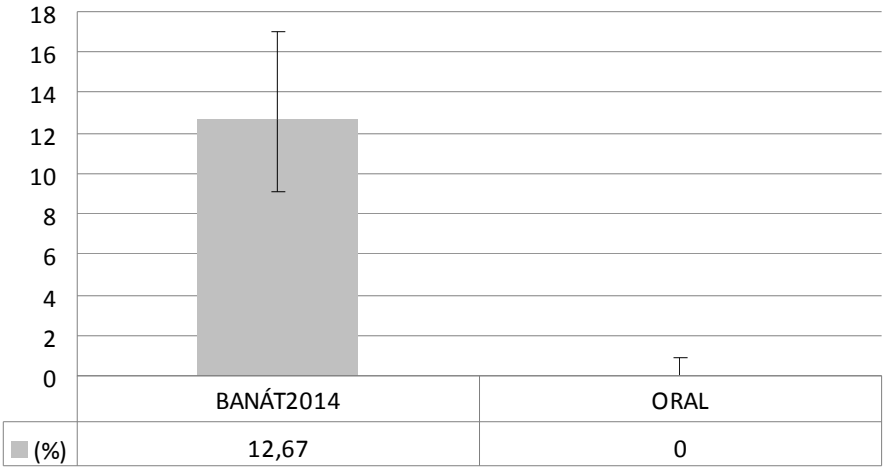
(34) *já sem si myslela že vona nebude letos dělat **nikerak** a hele vona dělá*
(35) *že a eee myslí že sme ale já nevim **nikerak***
(36) *bez šátku mně už to nesluší **nic***
(37) *žádnej to neuhasil . ani ten dobytek se nevysvobodil **nic***
(38) *neboj se nic tady hele --- ti to .. **nic** se neboj ..*
(39) *[vegetariánka] maso nejedla sádlo nejedla **nic** . jenom s volejem*
(40) *já se nebudu vysekávat **nic** . tak pudu jak sem .. jenom že si vemu šátek [...]*

Druhé zmíněné, zájmeno *nic*, by mělo funkci zesílení záporu plnit také (viz výše příklady (34)–(40), ostatní jsou v příloze 3). Tuto funkci však na rozdíl od *nikerak* neplní ve všech případech – obvykle vyjadřuje spíše *žádnou věc* anebo *něco v žádném množství* (SSJČ, heslo „nic“) – a bývá často spojeno se zdůrazňovací částicí *vůbec* (tj. spojení *vůbec nic*). Následující tabulka shrnuje výsledky korpusového dotazu:

		BANÁT2014	**ORAL**
počet výskytů hledaného lemmatu		468 (ipm 1 761,37)	1 318 (ipm 1 278,13)
velikost náhodného vzorku		300	400
z toho	***nic* („žádná věc“)**	262	400
	***nic* („vůbec“)**	38 (12,67 %)	0 (0 %)
konfidenční interval		9,122–16,97 %	0–0,9180%

Tabulka 12: Výsledky hledání *nic*

Fisherův test prokazuje statisticky významný rozdíl hodnotou *p-value* = 2,501e-15 < 0,05. *Nic* s funkcí zesíleného záporu je stejně jako v případě *nikerak* zcela charakteristickým pro bígerskou češtinu (viz graf 3).



Graf 4: Zesílený zápor vyjádřený zájmenem *nic*

6.2.2. Záporná zájmena a příslovce *nikam*, *nikde*, *nikdo*, *nikdy*, *nikoho* a *nikomu*

Společnými zápornými zájmeny a zájmennými příslovci v obou korpusech je těchto šest: *nikam*, *nikde*, *nikdo*, *nikdy*, *nikoho* a *nikomu* (výsledky jsou označeny tučně). V korpusu BANÁT2014 jsou navíc zachyceny výslovnostní varianty, které jsou vždy přiřazeny k danému lemmatu. Výsledky vzájemného srovnání shrnuje následující tabulka:

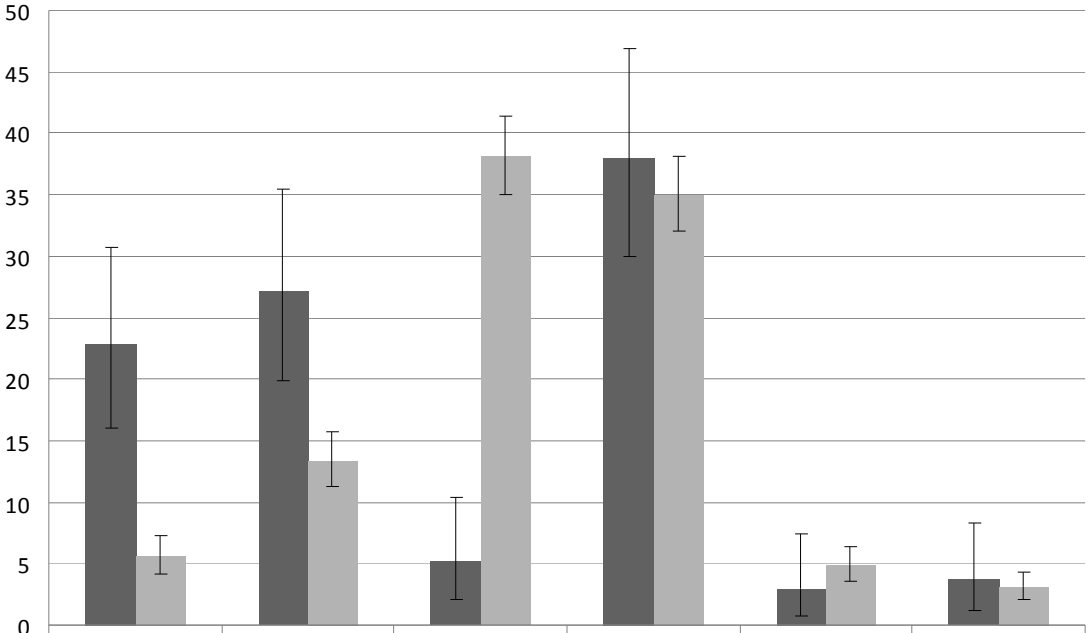
	BANÁT2014	**konf. (%)**	**ORAL**	**konf. (%)**	**p-value**
nikam	31 (22,79 %)	16,04–30,77	53 (5,573 %)	4,202–7,227	0,0000000015
nikde	37 (27,21 %)[42]	19,93–35,50	127 (13,35 %)	11,26–15,68	0,00009335
nikdo	7 (5,147 %)	2,094–10,32	363 (38,17 %)	35,07–41,34	< 2,2e-16
nikdy	52 (38 %)[43]	30,04–46,95	333 (35,02 %)	31,98–38,14	0,5024
nikoho	4 (2,941)	0,8071–7,359	46 (4,837 %)	3,563–6,399	0,5088
nikomu	5 (3,676)	1,204–8,371	29 (3,049 %)	2,052–4,350	0,604
celková populace	136 (imp 511,85)	-	951 (imp 922,23)	-	-

Tabulka 13: Srovnání *nikam*, *nikde*, *nikdo*, *nikdy*, *nikoho* a *nikomu*

[42] Započítáno i 8 tvarů *nikdež* a 1 *nikdeže* patřící pod stejné lemma *nikde*. Tyto varianty nebyly v korpusu ORAL zachyceny.

[43] Započítáno i 21 tvarů *nikdá* a 16 *nikdáž* patřící pod stejné lemma *nikde*. Tyto varianty nebyly v korpusu ORAL zachyceny.

Z výsledného srovnání zájmen a příslovcí *nikam*, *nikde*, *nikdo*, *nikdy*, *nikoho* a *nikomu* v případě *p-value* > 0,05 vyplývá, že užití *nikdy*, *nikoho* a *nikomu* nejsou signifikantně rozdílná v obou varietách češtiny. Naopak zájmena *nikam* a *nikde* jsou charakteristické pro bígerskou češtinu a zájmeno *nikdo* je signifikantně častěji užíváno v obecné češtině. Vše je znázorněno na následujícím grafu:



	nikam	nikde	nikdo	nikdy	nikoho	nikomu
BANÁT	22,79	27,21	5,147	38	2,941	3,676
ORAL	5,573	13,35	38,17	35,02	4,837	3,049

Graf 5: Srovnání *nikam, nikde, nikdo, nikdy, nikoho* a *nikomu*

6.3. Stálé příklonky *mi*, *ti*, *tě*, *si*, *se*, *mu*, *ho*

Tématem této podkapitoly je postavení stálých příklonek *mi*, *ti*, *tě*, *si*, *se*, *mu*, *ho*, tj. slov bez vlastního přízvuku, pro které je typické připojování do jednoho mluvního taktu s předcházejícím slovem. Problematika příklonek byla zpracována několika podrobnými studiemi, a to převážně na spisovné češtině (Esvan 2000; In-chon 2003; Toman 2000; Uhlířová 1987: 82–97, 2001).

Standardně jsou příklonky umístěny v tzv. druhé pozici. Druhou pozici je možno zjednodušeně charakterizovat následujícím způsobem (Uhlířová 1987: 83–84; zvýraznění K. V.): nachází se za libovolným větným členem (holým i rozvitým), za kterým může následovat

i vsuvka (41); za vedlejší větou (42); za spojkami s výjimkou *a*, *i*, *ale* (43). Druhá pozice může být obsazena jednou anebo i více příklonkami.

(41) *Hned v září **jsme se** sešli na první schůzce. Peněz, to bylo o něm všeobecně známo, **si** nevážil.*
(42) *Cestující, který nemá platný cestovní doklad, **se** vystavuje nebezpečí, že bude pokutován.*
(43) *Jana řekla, že **se** brzy vrátí. Už se Pavel rozhodl, nebo **si** to ještě rozmyslí?*

V mluveném jazyce je situace mnohem složitější, a to zejména proto, že syntax mluvených projevů je značně rozvolněná. Předně je někdy složité rozlišit začátek věty, od kterého se druhá pozice počítá, a to nejen v transkripci, ale i při poslechu. Někdy je situace jednoduchá, mluvčí se střídají téměř po replikách, poblíž příklonky se nachází spojka, podle které se velice dobře orientuje, nebo nějaké oslovení druhého mluvčího či podobné kontaktové výrazy, ale v jiných případech se jedná o dlouhé věty, které zdánlivě nikde nekončí nebo jsou naopak děleny pauzami na nelogických místech.

V některých případech začátku věty předchází pauza, ale problémem je, že ve značení často nemusí být poznat, kdy se jednalo o pouhé nadechnutí a kdy o delší zastavení, jediným řešením je pak konzultace se zvukem. V mnoha dalších případech je jednou tečkou naznačena pauza, ale reálně mluvčí pořád bezchybně navazuje na původní sdělení. Při poslechu by to ale nebylo označeno za situaci, kdy první promluvu přerušil a začal druhou. Proto jsem brala převážně ohled jen na pauzy typu „odmlčení“, a a případ od případu se musela rozhodovat, jestli byla větná struktura sdělení opravdu narušena. V mnoha případech komplikují analýzu výplňová slova (*jako*, *třeba*, *prostě*) anebo hezitační zvuky, na které jsem ohled nebrala, nesrozumitelné úseky, či překryvy mluvčích, které větu rozdělí do dvou částí, jež je třeba si pro analýzu znovu spojit.

Pro účely analýzy v této podkapitole jsem k rozlišování pozice, ve které se příklonka, nachází přistupovala tak, že základní je pozice druhá. Pro postavení příklonky na začátku věty používám termín „první pozice“, tj. příklonka se nachází na místě, které je před (byť hypotetickou) pozicí druhou. Situaci, kdy je příklonka kdekoliv za druhou pozicí nazývám „jinou pozicí“ (byť v příkladech je označena číslem tři).

Rozdělování a posuzování bylo časově poměrně náročné a patrně zdaleka ne dokonalé, protože třídění příklonek je mnohem složitější než rozlišování zájmena a slovesa v případě tvaru *má*. Občas nestačí výše popsaná pravidla a při anotaci je jedinou možností spolehnout se na intuici. Z toho důvodu by pro lepší popis bylo nejvhodnější porovnat sety dat mezi několika anotátory, ale to bohužel v rámci dílčí části diplomové práce není možné.

Prepozici zvratných zájmen *se* a *si* (tj. umístění na první a ne na obvyklou druhou pozici) jsem se věnovala ve dvou studiích (Vyskočilová 2012b: 51–53, 2013: 301–303).[44] Důležitý je až druhý zmíněný článek, který hypotézu, že je použití zvratných zájmen *se*, *si* na první pozici v bígerské češtině častější než v češtině na našem území (Viková (1994: 34) tvrdí, že vlivem rumunštiny), potvrdila v bígerské četině (p-value 0,002733, interval spolehlivosti 1,42–6,96 %, vyhledáváno bylo pomocí tagů, které formovaly podmínky, ve kterých se může hledané zájmeno nacházet). Nedošlo ale ke srovnání se stavem v obecné češtině.

Ostatní příklonky (*mi*, *ti*, *tě*, *mu*, *ho*), jejichž výčet byl opět čerpán z *Mluvnice současné češtiny* (Cvrček 2010: 57), byly zkoumány jen velice okrajově. Haiderová (2007: 90) na nich dokládá situace, kdy jsou příklonky umístěné až na konci promluvy (44)–(45), stejnou tendenci zachytil i Salzmann (1984: 105).

(44) *že bi koupil ešťe iní auto si*
(45) *kerej nebude poslouchat ho*

V následujících tabulkách 13–19 jsou stručně shrnuty poznatky o vybraných příklonkách. Způsob zkoumání byl u všech stejný. Pod každou tabulkou jsou uvedeny dva příklady (pro kompletní seznam příklonek mimo druhou pozici viz příloha 3). Vzhledem k výsledkům jsou všechna data okomentována pod tabulkou 20, která shrnuje výsledné *p-value* pro 1. a jinou pozici.

		BANÁT2014	**ORAL**
počet výskytů (populace)		91 (ipm 353,78)	2 366 (ipm 2 294,42)
velikost náhodného vzorku		-[45]	200
1. pozice		3 (3,297 %)	11 (5,5 %)
2. pozice		74	180
jiná pozice		6 (6,593 %)	7 (3,5 %)
nelze určit / chyba[46]		8	2
konfidenční interval	**1. pozice**	0,6851–9,333 %	2,777–9,628 %
	jiná pozice	2,458–13,8 %	1,419–7,078 %

Tabulka 14: Výsledky hledání příklonky *mi*

(46) *--- bolela tenkrát ani sem jim to neřekla **mi** řekli eště mam* maminko eště můžete ?*

[44] Hypotéza byla postavena na základě studií Haiderové (2007: 90), Skuliny (1978: 160) a Vikové (1994: 34).
[45] V takto označených případech byly ručně analyzovány všechny nalezené výskyty.
[46] V této kategorii jsou zařazeny situace, kdy příklonce předchází nesrozumitelný úsek nebo odmlčení, vedle kterého stojí příklonka osaměle, či případy, ve kterých je v transkripci nějaká chyba (např. *mi* místo *my*).

(47) *v sobotu a já dyž už viděli . že jako **mi** je lepšé tak mně povídali [...]*

		BANÁT2014	ORAL
počet výskytů (populace)		543 (ipm 2 043,64)	1 634 (ipm 1 584,57)
velikost náhodného vzorku		150	200
1. pozice		5 (3,333 %)	9 (4,5 %)
2. pozice		134	181
jiná pozice		7 (4,667 %)	7 (3,5 %)
nelze určit / chyba		4	3
konfidenční interval	**1. pozice**	1,091–7,607 %	2,078–8,370 %
	jiná pozice	1,897–9,79 %	1,419–7,078 %

Tabulka 15: Výsledky hledání příklonky *ti*

(48) *eee já sem jejich babička . sem **ti** to říkala*
(49) *dědku já už **ti** mám chuť plesknout .. jo jo ty si přilej eště*

		BANÁT2014	ORAL
počet výskytů (populace)		108 (ipm 406,47)	801 (ipm 776,77)
velikost náhodného vzorku		-	150
1. pozice		1 (0,9259 %)	1 (0,667 %)
2. pozice		106	149
jiná pozice		1 (0,9259 %)	0
nelze určit / chyba		0	0
konfidenční interval	**1. pozice**	0,02344–5,051 %	0,01688–3,658 %
	jiná pozice	0,02344–5,051 %	0–2,429 %

Tabulka 16: Výsledky hledání příklonky *tě*

(50) *ty seš furt uvázaná . **tě** bolí krk nebo copa ?*
(51) *jinak mě nic nenapadá (odmlčení) **tě** asi napadnu já*

		BANÁT2014	ORAL
počet výskytů (populace)		2 776 (ipm 10 447,76)	8 826 (ipm 8 558,99)
velikost náhodného vzorku		250	300
sloveso *být*		51	31
1. pozice		12 (4,8 %)	16 (5,333 %)
2. pozice		175	239
jiná pozice		10 (%)	6 (2 %)
nelze určit / chyba		2	8
konfidenční interval	**1. pozice**	2,504–8,234 %	3,079–8,517 %
	jiná pozice	1,935–7,233 %	0,7374–4,302 %

Tabulka 17: Výsledky hledání příklonky *si*

(52) ***si** zazvoníme [telefonem] . když chceme na tuty [vysílačky] bysme nemohli*
(53) *to je taky v pořádku . sem **si** tady moh dělat . ňáký takovýdle .. paličky*

		BANÁT2014	ORAL
počet výskytů (populace)		3 522 (ipm 13 255,40)	13 597 (ipm 13 185,65)
velikost náhodného vzorku		200	300
předložka *se*		5	4
1. pozice		10 (5 %)	13 (4,33 %)
2. pozice		181	92
jiná pozice		3 (1,5 %)	7 (2,333 %)
nelze určit / chyba		1	4
konfidenční interval	**1. pozice**	2,423–9,003 %	2,327–7,296 %
	jiná pozice	0,3104–4,321 %	0,9432–4,748 %

Tabulka 18: Výsledky hledání příklonky *se*

(54) *(odmlčení)* ***se*** *přišel podrbat co ?*
(55) *a jak ste tu tetu Marušku našli ? .. ste* ***se*** *ptali .. na ní ?*

		BANÁT2014	ORAL
počet výskytů (populace)		408 (ipm 1 535,55)	1 327 (ipm 1 286,85)
velikost náhodného vzorku		100	200
1. pozice		1 (1 %)	3 (1,5 %)
2. pozice		94	192
jiná pozice		3 (3 %)	5 (2,5 %)
nelze určit / chyba		2	0
konfidenční interval	**1. pozice**	0,02531–5,446%	0,3104–4,321 %
	jiná pozice	0,6230–8,518 %	0,8166–5,737 %

Tabulka 19: Výsledky hledání příklonky *mu*

(56) *takže jako mu to v* vynahradili tak ho .* ***mu*** *dali .. jednu kejtu*
(57) *no ale voni hned* ***mu*** *koupili druhou*

		BANÁT2014	ORAL
počet výskytů (populace)		769 (ipm 2 894,21)	1 720 (ipm 1 667,96)
velikost náhodného vzorku		150	200
1. pozice		2 (1,333 %)	2 (1 %)
2. pozice		136	196
jiná pozice		7 (4,667 %)	2 (1 %)
nelze určit / chyba		5	0
konfidenční interval	**1. pozice**	0,1619–4,733 %	0,1213–3,565 %
	jiná pozice	1,897–9,379 %	0,1213–3,565 %

Tabulka 20: Výsledky hledání příklonky *ho*

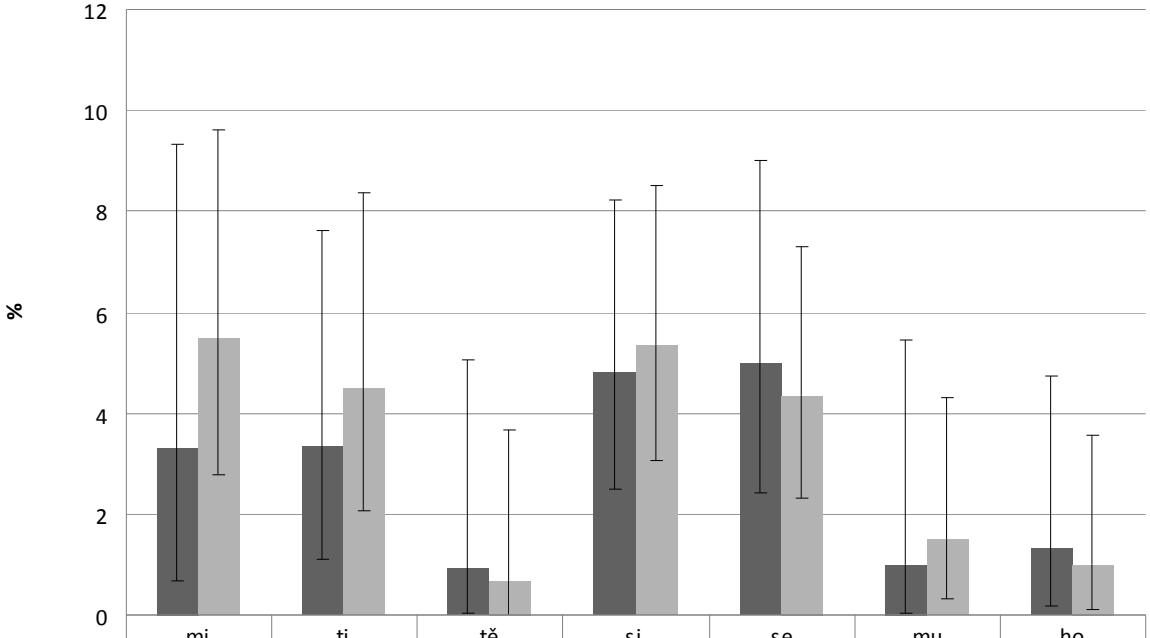
(58) *a vona prej* ***ho*** *nenašla to Inko [značka kávy]*
(59) *Jirka se tady učil německy měl slovníček sem* ***ho*** *zkoušela*

	mi	***ti***	***tě***	***si***	***se***	***mu***	***ho***
1. pozice	0,5601	0,7841	1	0,8471	0,8282	1	1
jiná pozice	0,2368	0,5935	0,4186	0,2051	0,7468	1	0,04169

Tabulka 21: *P-value* pro jednotlivé příklonky

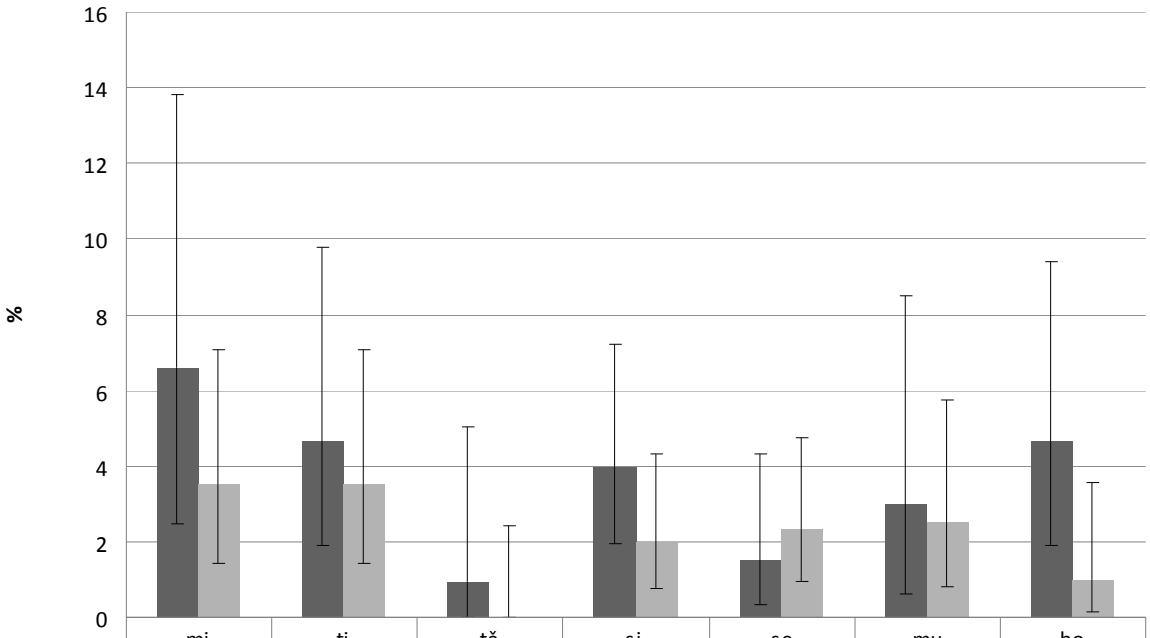
V užívání příklonek *mi*, *ti*, *tě*, *si*, *se*, *mu*, *ho* nebyla zjištěna odlišnost mezi banátskou češtinou a češtinou na našem území, to je dokázáno *p-value* > 0,05. Všechny hodnoty jsou uvedené v tabulce 20. Navíc se dle počtu výskytů jedná o řídký jev. Pro takové jevy je těžké získat signifikantní rozdíly, pokud je k dispozici pouze málo dat. Jak je vidět z konfidenčních intervalů znázorněných na naměřených hodnotách v následujících grafech, dané jevy se proto v korpusech ORAL a BANÁT2014 liší jen mírně. Pro přesnější výsledky by bylo nutné většího materiálu (např. u jiné pozice příklonky *mi* by mohl vyjít signifikantní rozdíl) a zároveň by bylo vhodné, aby značkování provádělo několik anotátorů a výsledná pozice zájmena tak byla výsledkem mezianotátorské shody.

Nesrovnalost s předchozím výzkumem zvratných zájmen *se* a *si* může být zapříčiněna tím, že příklonky byly vyhledávány pomocí tagů, nikoliv ručním značkováním.



	mi	ti	tě	si	se	mu	ho
BANÁT	3,297	3,333	0,9259	4,8	5	1	1,333
ORAL	5,5	4,5	0,667	5,333	4,33	1,5	1

Graf 6: Vybrané příklonky na první pozici



	mi	ti	tě	si	se	mu	ho
BANÁT	6,593	4,667	0,9259	4	1,5	3	4,667
ORAL	3,5	3,5	0	2	2,333	2,5	1

Graf 7: Vybrané příklonky na jiné než první nebo druhé pozici

7. ZÁVĚR

V samém úvodu práce byla stručně shrnuta historie Banátu a současný stav tamějšího českého osídlení. Tento úvod byl doplněn kapitolou věnovanou kontaktu banátské češtiny s češtinou spisovnou i obecnou (prostřednictvím mediální komunikace a školství) a také s cizími jazyky, zejména rumunštinou. Dále byl vytvořen přehled lingvistického výzkumu banátské češtiny od jeho počátku až po nejnovější práce.

Čtvrtá kapitola se věnuje tzv. specializovaným korpusům, mezi které patří i korpus BANÁT vytvářený v rámci této práce. Kromě definice specializovaných korpusů a komentáře k jejich budování byly součástí dvě podkapitoly, které mapují současnou situaci specializovaných mluvených korpusů českého jazyka a nabízí tak užitečný rozcestník k dalšímu jazykovému materiálu.

Těžištěm této práce byla tvorba specializovaného korpusu banátské češtiny BANÁT. Cílem bylo zachytit mluvenou češtinu mizejícího jazykového ostrova v Rumunsku a poskytnout materiál pro další možný výzkum. Zvolená folkloristická transkripce a další způsob zpracování dle korpusů řady ORAL projektu ČNK přinášejí možnost pohodlného srovnání bígerské jazykové variety a mluvené češtiny na našem území.

Sběr jazykového materiálu pro korpus proběhl v letech 2011 a 2014 v jedné ze šesti vesnic s kompaktním českým osídlením, v Bígru. Nahráno bylo více než 40 hodin spontánních rozhovorů osmnácti bígerských mluvčích a tří mluvčích českých. Způsob zpracování od surových dat až po vytvoření samotného korpusu byl detailně popsán a zdokumentován ve 4. kapitole. K dnešnímu dni obsahuje korpus BANÁT2014 přepisy 32,5 hodiny nahrávek a více než 210 000 textových slov bígerských mluvčích.

Na podzim 2014 je ve spolupráci s ÚČNK plánováno zpřístupnění finální verze korpusu BANÁT pro širší odbornou veřejnost. Finální korpus by měl obsahovat více než deset hodin dalších nahrávek, což s sebou přináší odhadem dalších 50 000 textových slov bígerských mluvčích. Vznikne tak dosud nejrozsáhlejší publikovaný soubor jazykového materiálu k banátské češtině. Hlavní výhoda korpusu BANÁT je ta, že bude propojen se zvukovou stopou, stejně jako ORAL2013. Budoucí uživatel si tak bude moci ve webových rozhraních NoSke a Kontext ke každému nalezenému segmentu promluvy pustit i jeho skutečnou zvukovou realizaci.

Do budoucna je možné korpus rozšiřovat o již existující jazykový materiál nasbíraný dalšími lingvisty v jiných vesnicích (Eibentál, Gerník, Svatá Helena a Šumice), v podobném časovém období a v podobné kvalitě.

V poslední části práce byl materiál obsažený v korpusu BANÁT2014 podroben lingvistické analýze vybraných jevů a srovnán se současným stavem obecné češtiny ve středočeské a jihozápadočeské nářeční oblasti, reprezentované daty z korpusu ORAL2013. Celý výzkum se věnoval především zájmenům, a to v první části přivlastňovacím, ve druhé záporným (doplněným i o zájmenné příslovce) a ve třetí pak zájmenným příklonkám.

Zkoumání přivlastňovacích zájmen potvrdilo hypotézu, že jsou zájmena *můj*, *jeho*, *její* a *jejich* užívána i v případech, kdy by mohlo být použito přivlastňovací zvratné zájmeno *svůj*. Frekvence tohoto jevu v bígerské češtině se signifikantně odlišuje od frekvence používání na našem území. V tomto případě se nejedná o pouhou statistickou signifikanci, z konfidenčních intervalů totiž vyplývá, že jsou zjištěné rozdíly opravdu lingvisticky významné. Také byl zjištěn signifikantní rozdíl ve frekvenci používání sledovaných zájmen jako takových.

V podkapitole věnované záporným zájmenům a příslovcím byl nejprve podroben analýze zápor zesílený pomocí zájmen *nic* a *nikerak*. Jev je běžně užíván v bígerské češtině, ale ve vybrané oblasti obecné češtiny nebyl doložen vůbec. Druhá část kapitoly se zabývala srovnáním zájmen a příslovcí *nikam*, *nikde*, *nikdo*, *nikdy*, *nikoho* a *nikomu* v obou varietách češtiny. Díky statistickému srovnání bylo zjištěno, že *nikdy*, *nikoho* a *nikomu* se v používání signifikantně neliší. Naopak příslovce *nikam* a *nikde* jsou charakteristická pro bígerskou češtinu a zájmeno *nikdo* pro češtinu na našem území.

Poslední podkapitola byla věnována poměrně náročnému srovnání příklonek *mi*, *ti*, *tě*, *si*, *se*, *mu*, *ho* a podrobná analýza nepotvrdila, že se tyto příklonky chovají v bígerské češtině odlišně od užití v češtině na našem území, jak předchozí výzkumy naznačovaly.

8. POUŽITÉ ZDROJE

8.1. Použité korpusy

BENEŠOVÁ, Lucie, Michal KŘEN a Martina WACLAWIČOVÁ, 2013. *ORAL2013: reprezentativní korpus neformální mluvené češtiny* [online]. Praha: Ústav Českého národního korpusu FF UK [cit. 2014–08–12]. Dostupné z: <http://www.korpus.cz>.

VYSKOČILOVÁ, Karolína, 2014. BANÁT2014: Korpus mluvené banátské češtiny [online]. Praha: Ústav Českého národního korpusu FF UK [cit. 2014–08–12]. Dostupné z: <http://www.korpus.cz>.

8.2. Odkazované webové stránky (bez autora)

About the Texas Czech Legacy Project [online] [cit. 2014-08-07]. Dostupné z: <http://blogs.utexas.edu/txczech/texas-czech-dialect-archive/>.

Akces: Akviziční korpusy českého jazyka [online] [cit. 2014-08-07]. Dostupné z: <http://akces.ff.cuni.cz/>.

aStat [online] [cit. 2014-08-13]. Dostupné z: <https://play.google.com/store/apps/details?id=org.twbbs.astat>.

Dostupné korpusy [online] [cit. 2014-08-07]. Dostupné z: <http://ucnk.ff.cuni.cz/struktura.php>.

Korpus DIALOG [online] [cit. 2014-08-12]. Dostupné z: <http://ujc.dialogy.cz/>.

Korpus Monolog [online] [cit. 2014-08-12]. Dostupné z: <http://monolog.dialogy.org/>.

PDTSL - Prague Dependency Treebank of Spoken Language [online] [cit. 2014-08-12]. Dostupné z: <http://ufal.mff.cuni.cz/pdtsl/>.

Post Bellum [online] [cit. 2014-08-12]. Dostupné z: <http://www.postbellum.cz/>.

Program podpory českého kulturního dědictví v zahraničí (krajané, lektoři) [online] [cit. 2014 h-07-16]. Dostupné z: <http://www.dzs.cz/cz/program-podpory-ceskeho-kulturniho-dedictvi-v-zahranici/>.

ROMi 1.0 [online] [cit. 2014 i-08-07]. Dostupné z: <http://ufal.mff.cuni.cz/romi-10>.

The Nijmegen Corpus of Casual Czech [online] [cit. 2014-08-13]. Dostupné z: <http://www.mirjamernestus.nl/Ernestus/NCCCz/>.

8.3. Citovaná literatura

ATKINS, Sue, Jeremy CLEAR a Nicholas OSTLER, 1991. Corpus Design Criteria. *Journal of Literary and Linguistic Computing* [online]. Oxford: Oxford University Press, č. 1, s. 1–16 [cit. 2014-05-27]. ISSN 0268-1145. Dostupné z: <http://www.natcorp.ox.ac.uk/archive/vault/tgaw02.pdf>.

BACHMANNOVÁ, Jarmila a Pavel JANČÁK, 2002. *Jak se mluvilo v českých vesnicích v cizině. Autentické zvukové ukázky z českých nářečí. Kladsko, Střelínsko, Žitomirsko, Daruvarsko, Banát*. Praha: Academia.

BALHAR, Jan aj., 1992–2005. *Český jazykový atlas 1–5*. Praha: Academia. ISBN 80-200- 0014-3, 80-200-0574-9, 80-200-0654-0, 80-200-0921-3, 80-200-1339-3.

BEDŘICHOVÁ, Zuzana, Karel ŠEBESTA, Svatava ŠKODOVÁ a Kateřina ŠORMOVÁ, 2011. Podoba a využití korpusu jinojazyčných a romských mluvčích češtiny: CZESL a ROMi. In: František ČERMÁK, ed. *Korpusová lingvistika Praha - 2 Výzkum a výstavba korpus*. Praha: Nakladatelství Lidové noviny, s. 93–104. ISBN 978-80-7422-115-6.

BEHRENS, Heike, 2008. *Corpora in language acquisition research: history, methods, perspectives*. Amsterdam: John Benjamins Pub. Co. ISBN 978-90-272-3476-6.

BENEŠOVÁ, Lucie, Michal KŘEN a Martina WACLAWIČOVÁ, 2013. *ORAL2013: reprezentativní korpus neformální mluvené češtiny* [online] [cit. 2014-08-12]. Dostupné z: <http://www.korpus.cz>.

BENEŠOVÁ, Lucie a Martina WACLAWIČOVÁ, 2014. Korpus neformální mluvené češtiny ORAL2013. In: Václav CVRČEK a Olga RICHTEROVÁ, ed. *Manuál práce s ČNK* [online]. [cit. 2014-07-17]. Dostupné z: <http://wiki.korpus.cz/doku.php?id=cnk:oral2013&rev=1391502113>.

BIBER, Douglas, 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing* [online]. Oxford: Oxford University Press, 1.10., roč. 8, č. 4, s. 243–257 [cit. 2014-07-14]. ISSN 0268-1145. Dostupné z: <http://staff.um.edu.mt/albert.gatt/teaching/dl/biber93.pdf>.

CIPLEA, Gheorghe, 1966. Influenţe româneşti în lexicul graiurilor cehe din Banat. *Cercetari de lingvistica*. roč. 11, s. 63–69. ISSN 0373-1545.

CIPLEA, Gheorghe, 1968. Consideratii privind elementele românesti in graiurile cehe din Banat. *Cercetari de lingvistica*. roč. 13, s. 237–244. ISSN 0373-1545.

CIPLEA, Gheorghe, 1971. Rumunské prvky v českých nářečích v rumunském Banátě. *Slavia*. Praha: Euroslavica, roč. 40, s. 211–219. ISSN 0037-6736.

COSTACHIE, Silviu, Elena BOGAN, Ionica SOARE a Adéla BARÁKOVÁ, 2011. Czech minority in Banat – Romania: A social geography survey. *Geographica Pannonica* [online]. roč. 15, č. 1, s. 7–15 [cit. 2014-06-25]. ISSN 1820-7138. Dostupné z: <http://www.dgt.uns.ac.rs/pannonica/papers/volume15_1_2.pdf>.

CVRČEK, Václav, 2010. *Mluvnice současné češtiny 1: Jak se píše a jak se mluví*. Praha: Karolinum. ISBN 978-80-246-1743-5.

ČERMÁK, František, nedatováno. *Pražský mluvený korpus* [online] [cit. 2014-08-12]. Dostupné z: <http://wiki.korpus.cz/doku.php/cnk:pmk>.

ČMEJRKOVÁ, Světla, 2003. Osudy zvratného posesivního zájmena svůj. *Naše řeč* [online]. roč. 86, č. 4, s. 181–205 [cit. 2014-08-09]. ISSN 0027-8203. Dostupné z: <http://nase-rec.ujc.cas.cz/archiv.php?art=7740>.

DOBRIŢOIU-ALEXANDRU, Teodora, 1965. Istoricul aşezării Cehilor în Banatul de sud (Republica socialistă România). *Romanoslavica (Filologie)*. Bukurešť: Editura Universităţii Bucureşti, roč. 12, s. 139–144.

DOBRIŢOIU-ALEXANDRU, Teodora, 1967. Banatismy v nářečích českých osad Svatá Helena, Gernik, Rovensko, Biger, Sumice a Clopodie. *Slavia*. Praha: Euroslavica, roč. 36, s. 374–382. ISSN 0027-8203.

ERNESTUS, Mirjam, Lucie KOČKOVÁ-AMORTOVÁ a Petr POLLAK, 2008. The Nijmegen Corpus of Casual Czech The Nijmegen Corpus of Casual Czech. In: *Proceedings of LREC 2014: 9th International Conference on Language Resources and Evaluation* [online]. s. 365–370 [cit. 2014-07-25]. Dostupné z: <http://www.mirjamernestus.nl/Ernestus/NCCCz/Ernestus_Ko%C4%8Dkova-Amortova_Pollak_2014_LREC.pdf>.

ESVAN, Francois, 2000. Česká klitika z hlediska typologického. In: Zeňka HLADKÁ a Petr KARLÍK, ed. *Čeština - univerzália a specifika 2*. Brno: FF MU, s. 141–148. ISBN 80-210-2262-0.

FISHER, Ronald Aylmer, 1922. On the Interpretation of $\chi 2$ from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society* [online]. roč. 85, č. 1, s. 87–94 [cit. 2014-07-29]. Dostupné z: <http://digital.library.adelaide.edu.au/dspace/bitstream/2440/15173/1/19.pdf>.

FRNOCHOVÁ, Adéla, 2012. *Jazyk české menšiny v obci Šumice v rumunském Banátě*. Praha. Diplomová práce, FF UK.

GESCE, Desideriu, 2013. *Historie českých komunit v Rumunsku*. Praha: Hermann & synové. ISBN 978-80-87054-31-4.

GOLÁŇOVÁ, Hana a Karina MATĚJŮ, 2008. Sociolingvistické aspekty koncepce Korpusu školní komunikace a Korpusu neformální komunikace dětí a mládeže. In: Marie KOPŘIOVOVÁ a Martina WACLAWIČOVÁ, ed. *Čeština v mluveném korpusu*. Praha: Nakladatelství Lidové noviny, s. 83–88. ISBN 978-80-7106-982-9.

GOLÁŇOVÁ, Hana a Karel ŠEBESTA, nedatováno. *Korpus vyučovacích hodin SCHOLA2010* [online] [cit. 2014-08-07]. Dostupné z: <http://ucnk.ff.cuni.cz/schola.php>.

GRANGER, Sylviane, 2002. A Bird's-eye view of learner corpus research. In: Sylviane GRANGER, Joseph HUNG a Stephanie PETCH-TYSON, ed. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam - Philadelphia: John Benjamins Publishing Company, s. 3–33. ISBN 1588112934.

HAIDEROVÁ, Karolina, 2007. *Jazyk české menšiny v rumunském Banátu: obce Gerník a Svatá Helena*. Olomouc. Diplomová práce, FF UPOL.

HAIDEROVÁ, Karolina, 2008. Archaismy v mluvě české menšiny v rumunském Banátu. In: Vladimír P. POLÁCH, ed. *Jazyková interakce a jazykové rozhraní a strategie „cutting-edge" [JazInt]. Sborník příspěvků z 8. mezinárodní konference Setkání mladých lingvistů konané na Filozofické fakultě Univerzity Palackého ve dnech 14.-16. května 2007.* Olomouc: Univerzita Palackého v Olomouci, s. 114–119.

HAIDEROVÁ, Karolina, 2010a. Jazyk mladé generace rumunských Čechů v Banátu. *Bohemica Olomucensia*. roč. 2, č. 3, s. 81–87.

HAIDEROVÁ, Karolina, 2010b. K vývojovým tendencím v deklinaci substantiv v současné rumunské češtině. In: *Minulost, přítomnost a budoucnost v jazyce a literatuře*. Ústí nad Labem: Univerzita Jana Evangelisty Purkyně v Ústí nad Labem, s. 358–362. ISBN 978-80-7414-362-5.

HAJIČ, Jan, Marie MIKULOVÁ, Martina OTRADOVCOVÁ, Petr PAJAS, Petr PODVESKÝ a Zdeňka UREŠOVÁ, 2006. *Pražský závislostní korpus mluvené češtiny. Rekonstrukce standardizovaného textu z mluvené řeči* [online]. [cit. 2014-08-02]. Dostupné z: <https://wiki.ufal.ms.mff.cuni.cz/_media/pdtsc:tr-2006-33.pdf>.

HAVRÁNEK, Bohumil, nedatováno. *Slovník spisovného jazyka českého* [online] [cit. 2014-08-12]. Dostupné z: <http://ssjc.ujc.cas.cz/>.

HLADKÁ, Zdeňka, nedatováno. *Brněnský mluvený korpus* [online] [cit. 2014-08-12]. Dostupné z: <http://wiki.korpus.cz/doku.php/cnk:bmk>.

HONOVÁ, P., 2000. *Čeština Čechů v rumunském Banátě - Rovensko*. Olomouc. Diplomová práce, PdF UP.

IN-CHON, Kim, 2003. *Fixed items in free word order languages: Clitics in Czech and sentence-final markers in Korean*. Dobřichovce: KAVA-PECH. ISBN 80-85853-70-1.

JECH, Jaromír, Milena SECKÁ, Vladimír SCHEUFLER a Olga SKALNÍKOVÁ, 1992. *Česká vesnice v rumunském Banátě*. Praha: Ústav pro etnologii a folkloristiku. ISBN 80-85010-36-4.

KARAS, František, 1937. *Československá větev, zapomenutá nebem i zemí I: Čechové v Rumunsku*. Praha: Spolek Komenský vlastním nákladem.

KOKAISL, Petr a KOL., 2009. *Krajané: po stopách Čechů ve východní Evropě*. Praha: Společnost „Za hranice". ISBN 9788025459249.

KOPŘIVOVÁ, Marie, Hana GOLÁŇOVÁ, Petra KLIMEŠOVÁ, Zuzana KOMRSKOVÁ a David LUKEŠ, v tisku. Multi-tier transcription of informal spoken Czech : the ORTOFON corpus approach. In: *Olomouc Modern Language Monographs*. Olomouc: Univerzita Palackého v Olomouci.

KOPŘIVOVÁ, Marie, Hana GOLÁŇOVÁ, Petra KLIMEŠOVÁ a David LUKEŠ, 2014. Mapping Diatopic and Diachronic Variation in Spoken Czech: the ORTOFON and DIALEKT Corpora. In: [online]. [cit. 2014-08-09]. Dostupné z: <http://www.lrec-conf.org/proceedings/lrec2014/pdf/252_Paper.pdf>.

KŘEN, Michal, 2013. *Odraz jazykových změn v synchronních korpusech*. Praha: Nakladatelství Lidové noviny. ISBN 9788074222658.

KŘIVAN, Jan a Magdalena ZÍKOVÁ, 2014. Nahrávání v terénním lingvistickém výzkumu: jak získat kvalitní záznam řeči? *Studie z aplikované lingvistiky / Studies in Applied Linguistics*. Praha: Univerzita Karlova v Praze, Filozofická fakulta, roč. 1, s. 65–82. ISSN 1804–3240.

LAMPRECHT, Arnošt, 1976. *České nářeční texty*. Praha: SPN.

LAZU, Vlasta, 2010. Historie českého vysílání Rumunského rozhlasu. *Český dialog* [online]. roč. 3-4 [cit. 2014-07-27]. Dostupné z: <http://www.cesky-dialog.net/clanek/4418-historie-ceskeho-vysilani-rumunskeho-rozhlasu/>.

MCENERY, Tony a Andrew HARDIE, 2012. What is corpus linguistics? In: *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press, s. 1–23. ISBN 978-0-521-54736-9.

MILIČKA, Jiří, 2012. Rank-frequency Relation & Type-token Relation: Two Sides of the Same Coin [online]. s. 1–11 [cit. 2014-08-07]. Dostupné z: <http://milicka.cz/kestazeni/type-token_beograd_2012.pdf>.

MILIČKA, Jiří, nedatováno. *Lexicographers' Calculator* [online] [cit. 2014-08-13]. Dostupné z: <http://milicka.cz/lexicographerscalculator/>.

MILIČKA, Jiří, připravuje se. Konfidenční intervaly v empirické lingvistice.

MORAVCOVÁ, Ilona, 2006. České menšinové školy v rumunském Banátě – od jejich založení do vypuknutí 2. světové války. In: *Theatrum historiae : sborník prací Katedry historických věd Fakulty filozofické Univerzity Pardubice* [online]. Pardubice: Univerzita Pardubice 1, s. 281–308 [cit. 2014-08-04]. Dostupné z: <http://dspace.upce.cz/handle/10195/35045>.

ORÁLKOVÁ, Helena, 2013. *Vybrané kapitoly z onomastiky a dialektologie vesnice Gerník (rumunský Banát)* [online]. Brno [cit. 2014-07-25]. Diplomová práce, FF MU. Dostupné z: <http://is.muni.cz/th/217507/ff_m/>.

PLETER, Tiberiu, 1965. Verbul în graiurile Cehe din Banat. *Analele Universitaţii Bucureşti: Seria Ştiinţe sociale (Filologie)*. roč. 14, s. 369–375.

POŘÍZKA, Petr, 2007. Olomoucký mluvený korpus – stav, metodologie, charakteristika. In: František ŠTÍCHA a Mirjam FRIED, ed. *Grammar adn Corpora / Gramatika a korpus*. Praha: Academia, s. 191–198.

ROHÁĽOVÁ, Kateřina, 2013. Závěrečná zpráva 2012/2013 - Eibenthal [online]. [cit. 2014-07-18]. Dostupné z: <http://www.dzs.cz/file/577/2012-13, Rumunsko, Eibenthal.pdf>.

SALZMANN, Zdeněk, 1983. *Two contributions to the study of Czechs and Slovaks settled in Romania*. Occasional. Amherst: University of Massachusetts of Amherst.

SALZMANN, Zdeněk, 1984. Some observations on the Czech spoken by the villagers of Ravensca in the Southern Romanin Banat. *Melbourne Slavonic Studies*. roč. 18, s. 65–118.

SALZMANN, Zdeněk, 1993. Přezdívky v Bigăru: příspěvek k antoponymii česky mluvící vesnice v jihorumunském Banátu. *Naše řeč* [online]. roč. 76, č. 3 [cit. 2014-07-29]. ISSN 0027-8203. Dostupné z: <http://nase-rec.ujc.cas.cz/archiv.php?art=7140>.

SALZMANN, Zdeněk, 2004. Bibliografie prací o rumunských Češích a Slovácích. *Češi v cizině*. Praha: Etnologický ústav AV ČR, roč. 12, s. 144–176.

SECKÁ, Milena, 1995. Češi v rumunském Banátu. *Češi v cizině*. Praha: Etnologický ústav AV ČR, roč. 9, s. 92–115. ISSN 0027-8203.

SCHLÖGL, Jindřich, 1925. *Dějiny českých osad v rumunském banátě*. Praha: Národní rada československá.

SINCLAIRE, John, 1996. Preliminary recommendations on Corpus Typology [online]. [cit. 2014-08-01]. Dostupné z: <http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html>.

SKARNITZL, Radek, 2010. Prague Phonetic Corpus: status report. *AUC Philologica 1/2009, Phonetica Pragensia XII*. s. 65–67. ISSN 0567-8269.

SKOŘEPA, Petr, 2013a. *Závěrečná zpráva 2012/13 - Gerník* [online]. [cit. 2014-07-18]. Dostupné z: <http://www.dzs.cz/file/581/2012-2013, Rumunsko, Gern%C3%ADk.pdf>.

SKOŘEPA, Petr, 2013b. *Závěrečná zpráva 2012/13 - Svatá Helena* [online]. [cit. 2014-07-18]. Dostupné z: <http://www.dzs.cz/file/582/2012-2013, Rumunsko, Svat%C3%A1 Helena.pdf>.

SKULINA, Josef, 1974. Zdvojené p ředložky do + na , na + pod , kolem + do , pod + za v českém nářečí na území rumunského Banátu. *Naše řeč* [online]. roč. 57, č. 3 [cit. 2011-04-16]. ISSN 0027-8203. Dostupné z: <http://nase-rec.ujc.cas.cz/archiv.php?art=5774>.

SKULINA, Josef, 1975. Banatismy v češtině na území Rumunska. *Sborník prací FF Brněnské univerzity, Series linguistica (A)*. Brno: Univerzita J. E. Purkyně v Brně - filozofická fakulta, roč. 23-24, s. 69–73.

SKULINA, Josef, 1978. Banátská čeština. *Sborník prací FF Brněnské univerzity, Series linguistica (A)*. Brno: Univerzita J. E. Purkyně v Brně - filozofická fakulta, roč. 26-27, s. 157–163.

SKULINA, Josef, 1979. Zvláštnosti bilingvismu rumunských čechů. *Zborník FF Univerzity Komenského, Philologica 30*. s. 227–232.

ŠEBESTA, Karel, 2010. Korpusy češtiny a osvojování jazyka. *Studie z aplikované lingvistiky / Studies in Applied Linguistics*. č. 2, s. 11–34. ISSN 1804-3240.

ŠEBESTA, Karel a Svatava ŠKODOVÁ, 2012. *Čeština - cílový jazyk a korpusy*. Liberec: Technická univerzita v Liberci. ISBN 978-80-737-2848-9.

TOMAN, Jindřich, 2000. Prosodické spekulace o klitikách v nekanonických pozicích. In: Zdeňka HLADKÁ a Petr KARLÍK, ed. *Čeština - univerzália a specifika 2*. Brno: FF MU, s. 161–166. ISBN 80-210-2262-0.

TŮMOVÁ, Markéta, 2011. *Jazyk rumunských Čechů reemigrantů - současný stav* [online]. Praha [cit. 2014-07-25]. Bakalářská práce, PedF UK. Dostupné z: <https://is.cuni.cz/webapps/zzp/detail/95642/>.

UHLÍŘOVÁ, Ludmila, 1987. *Knížka o slovosledu*. Academia.

UHLÍŘOVÁ, Ludmila, 2001. Klitika jako kategorie diskusní. In: Zdeňka HLADKÁ a Petr KARLÍK, ed. *Čeština - univerzália a specifika 3*. Brno: FF MU, s. 29–35. ISBN 80-210-2532-8.

URBAN, Rudolf, 1930. *Čechoslováci v Rumunsku*. 1930. Bukurešť: Československý ústav zahraniční.

UTĚŠENÝ, Slavomír, 1962. O jazyce českých osad na jihu rumunského Banátu. *Český lid* [online]. roč. 49, č. 5, s. 201–209 [cit. 2014-07-15]. ISSN 0009-0794. Dostupné z: <http://www.svata-helena.eu/hs-download/historie/Cesky_lid_5-1962.pdf>.

UTĚŠENÝ, Slavomír, 1964a. Vlastní jména osob a zví řat u Čechů na jihu rumunského Banátu. *Naše řeč* [online]. roč. 47, č. 4 [cit. 2014-07-15]. ISSN 0027-8203. Dostupné z: <http://nase-rec.ujc.cas.cz/archiv.php?art=5060>.

UTĚŠENÝ, Slavomír, 1964b. Z druhé výpravy za češtinou v rumunském Banátě. *Český lid*. roč. 51, s. 27–32. ISSN 0009-0794.

UTĚŠENÝ, Slavomír, 1970. O posrbšťování kruščické češtiny v jugoslávském Banátě. *Naše řeč* [online]. roč. 53, č. 3 [cit. 2014-07-16]. ISSN 0027-8203. Dostupné z: <http://nase-rec.ujc.cas.cz/archiv.php?art=5496>.

VAŠEK, Antonín, 1968. K vzájemným vztahům slovanských a neslovanských jazyků v rumunském Banátě. In: *Československé přednášky pro VI. meznárodní sjezd slavistů*. Academia. s. 165–169.

VAŠEK, Antonín, 1975. K vývoji „izolovaného“ slovanského jazyka. *Slavia*. Praha: Euroslavica, roč. 44, s. 1–6. ISSN 0027-8203.

VIKOVÁ, Vilma, 1994. *Bígerská čeština: nástin jazykové monografie české vesnice v Rumunsku*. Praha. Diplomová práce, FF UK.

VOLÍN, Jan, Lenka WEINGARTOVÁ a Oliver NIEBUHR, 2014. Between Recognition and Resignation – The Prosodic Forms and Communicative Functions of the Czech Confirmation Tag “ jasně ” Institute of

Phonetics , Charles University in Prague , Czech Republic. In: N. CAMPBELL, D. GIBBON a D. HIRST, ed. *Proceedings of the 7th International Conference on Speech Prosody*. Dublin: TCD, s. 115–119.

VOLÍN, Jan a Lenka WEINGARTOVÁ, 2014. Současný stav zkoumání zvukové stránky mluvních stylů.

VYSKOČILOVÁ, Karolína, 2012a. Korpus BANÁT. *Tvar*. roč. 17, č. 12, s. 10.

VYSKOČILOVÁ, Karolína, 2012b. *Syntaktická analýza projevů českých mluvčích v rumunském Banátu*. Praha. Bakalářská práce, FF UK.

VYSKOČILOVÁ, Karolína, 2013. K slovosledu mluvené češtiny v rumunském Bígru. In: Božena BEDNAŘÍKOVÁ a Pavla HERNANDEZOVÁ, ed. *Od slova k modelu jazyka: Sborník z 13. mezinárodního setkání mladých lingvistů*. Olomouc: Univerzita Palackého v Olomouci, s. 296–309. ISBN 978-80-244-3960-0.

VYSKOČILOVÁ, Karolína, v tisku. Czech language minority in the South Western Romanian Banat. *Multilingualism and Minorities in the Czech Sociolinguistic Space. International Journal of the Sociology of Language*.

WACLAWIČOVÁ, Martina, Marie KOPŘIVOVÁ a Olga RICHTEROVÁ, nedatováno. *Český mluvený korpus ORAL2006* [online] [cit. 2014-08-12]. Dostupné z: <http://wiki.korpus.cz/doku.php/cnk:oral2006>.

WACLAWIČOVÁ, Martina a Olga RICHTEROVÁ, nedatováno. *Český mluvený korpus ORAL2008* [online] [cit. 2014-08-12]. Dostupné z: <http://wiki.korpus.cz/doku.php/cnk:oral2008>.

WALLIS, Sean, 2013. Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. *Journal of Quantitative Linguistics*. roč. 20, č. 3, s. 178–208. ISSN 0929-6174.

WEINGARTOVÁ, Lenka, Eliška CHURAŇOVÁ a Pavel ŠTURM, 2014. *Transitions, pauses and overlaps: Temporal characteristics of turn - taking in Czech*. Dublin: TCD.