Korpusová lingvistika: Stav a modelové přístupy

prof. PhDr. František Čermák, DrSc. (ed.)
PhDr. Renata Blatná, CSc. (ed.)

NAKLADATELSTVÍ LIDOVÉ NOVINY

Ústav Českého národního korpusu

někdy tak sedím a **<chce se mi>** *vám všem třeba napsat dopis nebo zavolat , ale pak si říkám*
čtenář chápavý , ale **<chce se mi>** *věřit , že jsi vyjádřil asi tyto mé pocity : – Svoboda , kterou*
být v pořádku a **<chce se mi>** *věřit , že nová sezóna bude lepší , než ta minulá , dost*
neúřední jazyk , ale **<chce se mi>** *z toho blít ." „Dostane ho zpátky ? „ Přikývl .*
podívat a stále **<chce se mi>** *zaplakat Nad tebou nade vším co zdělilo tě nad ní kterou*
když já jej ucítím, **<chce se mi>** *zaplakati . Zvolna se otáčí , jak by se ohlížel , jak by se*
krásných stromů **<chce se mi>** *zas jednou domů jít , musí to tak být . Mamuti Nezmaři*
se kolem sebe a **<chce se mi>** *zase začít láteřit , že nejsou peníze na kulturu . Jenomže ony*
se všemi jeho dny, **<chce se mi>** *zemřít a nebýt trpělivý a skončit to zlé čekání ! Nenávidím*
MS . Pokud ne, **<chce se mi>** *zvolat , i když nejsem věřící , „ Bůh nás chraň „ . Aleš*
kdy mám depresi a **<chce se mi>** *zvracet , protože jsem snědl něco , co nebylo zrovna*
„ řekl Donnelly , „ **<chce se mi>** *zůstat a opít se ." „Než půjdete spát , vezměte si dva aspiriny*

Srovnávací rozbor mluvených korpusů (PMK a BMK): metodologické problémy a první výsledky

François Esvan

1. Úvod

Bádání o městských mluvách se od diskuse o obecné češtině v polovině šedesátých let považuje za hlavní úkol české sociolingvistiky. Během tohoto období se uskutečnilo víc průzkumů různých částí národního území (viz retrospektiva v nedávném článku od P. Sgalla 2004). Co se týče Prahy nebo středních Čech, jde hlavně o rozbory korpusů nahraných mluvených projevů, mezi nimiž nejvýznamnější jsou následující:

– Kravčišinová et al. 968: mluvený korpus o rozsahu 10 tisíc vět z každodenního hovoru a z rozhlasových pořadů v období 1962–63 (celkově 79 mluvčích pocházejících z Čech, různého věku, profese a pohlaví);
– Hammer 1985: nahrávky 29 osob narozených anebo žijících v Praze víc než 25 let. Výběr je omezený na vzdělané lidi ve věku od 35 do 55 let (viz o tom Sgall et al. 1992, 192);
– Šonková 2000: součást PMK o rozsahu 40 tisíc slov (50 mluvčích různého věku, pohlaví a vzdělání);
– Maglione 2003: 118 stran soukromých rozhovorů, televizních pořadů (1999–2000) a debat Občanského fóra (1989). Celkově 27 mluvčích.

V případě Brna je situace podstatně jiná, protože disponujeme různými studiemi od stejné autorky (Krčmová 1981, 1997), které jsou naopak založené na metodě dotazníků. Na konci devadesátých let se otevřely nové perspektivy s budováním dvou volně dostupných mluvených korpusů v Praze a v Brně (PMK a BMK). Přestože mají daleko větší rozsah a jsou mnohem lépe strukturované než všechny dosud používané korpusy (BMK: 500 tisíc slov, 250 nahrávek od 294 mluvčích v období 1994–1999; PMK: 675 tisíc slov, 304 nahrávek v období 1988–1996), těmto zdrojům zatím nebyla věnována zasloužená pozornost (viz jenom některé sondy, např. Kopřivová 2004).

Výhoda PMK a BMK tkví mj. v tom, že oba korpusy byly vytvořeny přibližně ve stejném období a vycházejí z téže koncepce (Hladká 2005). Lze tedy uvažovat o srovnávacím rozboru, což je v českém sociolingvistickém kontextu obzvlášť zajímavé kvůli tomu, že tzv. pronikání obecné češtiny je na Moravě centrálním bodem diskuse. Toto pronikání, jímž se M. Krčmová zabývá ve svých pracích

(1997, 227–228) již řadu let, lze při srovnání mezi BMK a PMK poprvé systematicky pozorovat na základě korpusových dat.

Než začneme se srovnávacím rozborem BMK a PMK, je však potřeba probrat několik metodologických problémů, jimiž jsou: povaha a strukturace korpusů (2.); typy údajů (3.); statistická platnost získaných údajů (4). K těmto metodologickým úvahám jsme si vybrali jako názornou ukázku případ kolísání mezi adjektivními koncovkami *-ý* a *-ej*. Tento případ bude v druhé části co nejúplněji rozebrán na základě různých typů měření, nejdříve podle jednotlivých lexémů (5.) a potom globálně (6.). Všechny výsledky budou následně vyhodnoceny z hlediska statistické významnosti a interpretovány podle jednotlivých parametrů (věk, pohlaví, vzdělání a formálnost) (7.)

2. Povaha a strukturace korpusů

PMK a BMK jsou cennými sbírkami nahrávek mluvených hovorů, které lze používat k nejrůznějším cílům. Je například možné z nich čerpat autentické příklady ke studiu nejrůznějších lingvistických jevů, od syntaxe po morfologii a lexikon. K tomuto účelu se obyčejně používá funkce vyhledávání konkordancí a zjištěné údaje jsou pak analyzovány z hlediska kvalitativního. Výhoda korpusových dat tkví v tom, že získané příklady pocházejí z velkého počtu zdrojů, a ne jenom z intuice pozorovatele. Pokud je užívaný korpus reprezentativní, je pak možné zjištěné výsledky inferovat a prohlásit, že neplatí jen pro daný vzorek, ale všeobecně (zatím však lze jen těžko mluvit o reprezentativních mluvených korpusech, viz Čermák 2006).

Rozbor, který zde uvádíme, však patří do zcela jiné kategorie, protože nevychází z konkrétních příkladů, ale pouze z kvantitativních údajů. Pokud se vezme korpus jako zdroj příkladů, základní data jsou věty a každá věta je apriorně zajímavá, protože je sama sobě konfrontačním prvkem s vlastní intuicí pozorovatele. Číselné údaje takovou podstatnou platnost vůbec nemají. Musí se podřídit přesným pravidlům, která jsou základem jakéhokoliv vědeckého zkoumání založeného na statistikách. Tato pravidla se týkají především povahy vzorku, na jehož základě se provádí měření. Jak bylo řečeno, PMK a BMK mají větší spektrum užívání a nebyly vytvářeny jenom kvůli kvantitativním rozborům. Je proto potřeba, aby se předem zjistilo, za jakých podmínek se k tomuto účelu dají používat.

Podle teorie statistiky lze nějaký vzorek považovat za reprezentativní pro celou populaci jedině v případě, že byl vytvořen náhodně. Náhodná metoda není však z různých důvodů v sociolingvistice představitelná a používají se místo toho tzv. stratifikované vzorky (viz např. Davis 1990, 1–11). Prakticky se předem rozhoduje, jaké sociologické parametry jsou pro dané studium významné a korpus se potom buduje takovým způsobem, aby byl každý parametr reprezentován. Aby bylo možné získané výsledky inferovat, musí být korpus zároveň reprezentativní, což znamená, že dotyčné parametry by měly

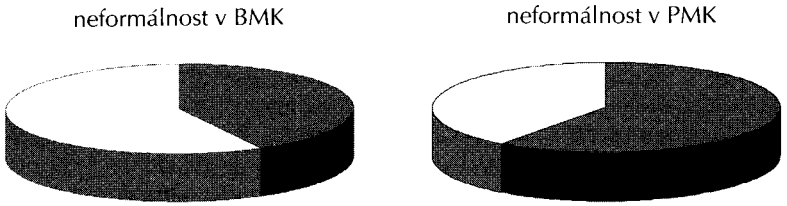
být v teorii ve stejné proporci jako v celkové populaci (např. stejný počet mladých, žen, vzdělaných lidí apod.).

Lze taky vytvářet vzorky za účelem zjištění, zda nějaký lingvistický jev je ovlivněn určitými parametry. V takovém případě je důležité počítat s možnou interferencí jiných parametrů a vyhýbat se tzv. destruktivním distorzím. Je-li například cílem průzkumu zjistit, zda nějaký jev je běžnější v písemném anebo v mluveném jazyce, je velmi důležité vycházet z mluvených projevů, které nejsou jenom formální, protože formálnost je destruktivním faktorem v tom smyslu, že formální projevy mívají rysy psanosti. Taková metoda je ovšem dost aproximativní, protože je jasné, že se berou v úvahu jenom ty jevy, o kterých se už něco předem ví anebo tuší. Pokud se však aplikuje s dostatečnou přesností, lze považovat získané výsledky za dostatečně spolehlivé.

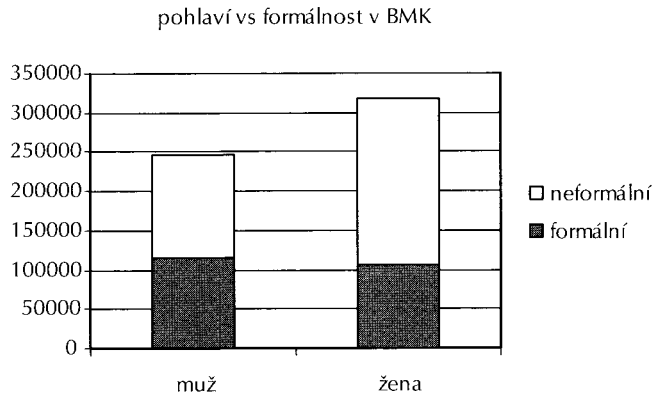
Jak to tedy vypadá z tohoto hlediska s BMK a PMK? Oba korpusy byly vytvořeny na základě sociolingvistické stratifikace podle čtyř klasických parametrů, jimiž jsou věk, pohlaví, vzdělání a formálnost (další informace o tom lze získat na internetových stránkách ÚČNK). Tvůrci BMK a PMK dbali, aby všechny možné kombinace těchto parametrů byly v korpusech přítomné (je jich celkem 2^4=16: IUNIOR + MUŽ + NEFORMÁLNÍ + BASIS, IUNIOR + MUŽ + NEFORMÁLNÍ + ALTUS atd.). Netrvali však na tom, aby byla dodržována přesná rovnováha mezi různými subkategoriemi. Počet slov v každé subkategorii je označen v následující tabulce:

	BMK		**PMK**	
	N	%	N	%
altus	324371	54,9	371218	46
basis	266851	45,1	436655	54
formální	250282	42,3	487479	60,3
neformální	340940	57,7	320394	39,7
iunior	387409	65,5	423635	52,4
vetus	203813	34,5	384238	47,6
muž	246086	41,4	365233	45,2
žena	348136	58,6	442730	54,8
Směrodatná odchylka		10,8		6,6

Lze konstatovat, že kvantitativní rozdíly mezi subkategoriemi nejsou zdaleka zanedbatelné. Například poměr v procentech mezi subkategoriemi FORMÁLNÍ – NEFORMÁLNÍ je v BMK 42,3 : 57,7 a v PMK 60,3 : 39,7, což znamená, že PMK je bezesporu „formálnější" než BMK a že při přímém srovnání musí dojít k chybným výsledkům.



Tato nerovnováha je překážkou nejen pro srovnávací rozbor mezi korpusy, ale i pro analýzu významnosti parametrů **uvnitř stejného korpusu**. Problém lze znázornit konkrétním příkladem. Připustí-li se, že formálnost působí na užívání koncovky *-ej*, není pochyb, že příliš velký počet formálních projevů v subkategorii ŽENA může vést ke klamné domněnce, že pohlaví má významný vliv:



Jediný způsob, jak čelit těmto nedostatkům, je provést korekci u všech dotyčných faktorů. Vzhledem k tomu, že korpusy jsou stratifikovány na základě čtyř parametrů, je potřeba tuto korekci vykonat v 16 subkategoriích podle počtu možných kombinací (viz výše). Prakticky se tedy musí zaznamenat počet výskytů v každé subkategorii zvlášť a korigovat výsledky podle celkového počtu slov dotyčné subkategorie. Korigovaná frekvence Fki se počítá podle následujícího vzorce:

korigovaná frekvence = (změřená frekvence / počet slov v subkategorii) x koeficient korekce

Koeficient korekce je spočítán tak, aby se vztahoval v každém korpusu k normalizované subkategorii průměrné velikosti rovnající se celkovému počtu slov děleného počtem subkategorií (BMK 591 221 : 16 = 36 951, PMK 807 873 : 16 = 50 492.). Na základě těchto korigovaných hodnot lze spočítat korigovanou globální frekvenci v celém korpusu (součtem všech hodnot), a korigované frekvence pro každý jednotlivý parametr (součtem různých kombinací hodnot).

Kvantitativní nerovnováha mezi subkategoriemi nás tedy nutí k pracnému systému měření ve všech subkategoriích, ale tím lze právě získat všechny potřebné údaje k nejrůznějším rozborům, globálním anebo rozděleným podle jednotlivých parametrů (viz níže 7).

Rozdělení sčítání výskytů podle subkategorií zvyšuje množství práce a značně komplikuje rozbor údajů, což je nevýhoda z čistě praktického hlediska. Toto rozdělení však působí negativně i na jiné faktory. Dotyčné subkategorie obsahují totiž zmenšený počet slov (přesně 1/16 celkového počtu), což omezuje možnosti zkoumání na jevy s relativně vysokou frekvencí. Je-li totiž frekvence příliš nízká, počet výskytů se v jednotlivých subkategoriích stane minimálním nebo dokonce často nulovým. Korekce ovšem ztratí v takových případech význam, a nabízenou metodu už nelze používat (o tomto problému při rozboru podle jednotlivých lexémů viz níže).

Ve skutečnosti je korekce pouze východisko z nouze, jímž se snažíme kompenzovat relativní nevhodnost vzorků pro tento typ kvantitativního rozboru. Bylo by samozřejmě lepší disponovat stratifikovanými korpusy, ve kterých by měla každá subkategorie stejný rozměr. Tento problém se však nedá vyřešit tak jednoduše, protože rovnováha vyžaduje nivelizaci rozsahu v každé subkategorii podle nejmenšího počtu slov, což není bez vlivu na celkovou velikost korpusu: J. Šonková například normalizovala ve své studii počet slov, aby měly soubory stejný rozměr pro všechny mluvčí, a z toho důvodu se jí rozsah pracovního korpusu zmenšil z 127 tisíc na pouhých 47 tisíc slov (Šonková 2000, 192).

3. Typy údajů

3.1. Typologie

Existují různé způsoby, jak kvantitativně hodnotit lingvistické jevy. Při letmém pohledu na existující korpusové studie vyjde najevo, že o metodologických problémech se příliš nemluví a že pracnost získání dat je při výběru přístupu důležitější než otázka jejich významnosti. Pokusíme se zde na tyto metodologické problémy podívat trochu podrobněji na příkladu kolísání mezi adjektivními koncovkami *-ej* a *-ý*. Pro takové jevy je možné všeobecně uvažovat o různých kombinacích mezi následujícími základními přístupy:

A) měření absolutní vs relativní: hodnoty lze měřit **absolutně** (spočítáním tvarů na *-ej*), anebo **relativně** (spočítáním tvarů na *-ej* a zároveň na *-ý* a hodnocením v procentech).
B) počítání individuální vs globální: výskyty lze počítat **individuálně** pro každý lexém, anebo **globálně** pro všechny lexémy dohromady (v našem případě pro všechna adjektiva typu *hezký*).

V následujících odstavcích rozebereme výhody a nevýhody měření absolutního (3.2.), relativního (3.2.) a potom globálního počítání ve srovnání s individuálním (3.3.).

3.2. Měření absolutní

Nevýhoda absolutního měření spočívá v tom, že se hodnoty dají interpretovat jedině s ohledem na velikost referenčního korpusu. K tomu, aby se daly srovnávat absolutní hodnoty získané v korpusech různých velikostí, je tedy potřeba provést nějakou korekci. M. Kopřivová (2004, 169) srovnává například hodnoty získané v BMK, PMK, BELETRIE a SYN2000 (mimo jiné bere v úvahu i případ koncovky *-ej*) po normalizaci na základě konvenčního korpusu velikosti 100 miliónů slov.

Jak bylo řečeno výše, kvantitativní nerovnováha mezi subkategoriemi nutí k systematické korekci, která se týká všech hodnot jak při měření absolutním, tak relativním. Chtěli bychom teď upozornit na jiný zdroj distorze, který není tak triviální a působí jenom v případě absolutního měření. Tato distorze vyplývá z toho, že se lexémy anebo vůbec gramatické kategorie mohou vyskytovat více anebo méně často podle takových sociologických parametrů, jako jsou pohlaví, vzdělaní, věk anebo formálnost. Podívejme se na výsledky označené v následující tabulce, které se týkají koncovek *-ej*. Jde se o korigované a nekorigované absolutní hodnoty podle kategorie FORMÁLNÍ – NEFORMÁLNÍ (N – F) v případě čtyř různých lexémů, *takový, nějaký, každý* a *který*:

	takovej				***nějakej***			
	nekorig.		korig.		nekorig.		korig.	
	BMK	PMK	BMK	PMK	BMK	PMK	BMK	PMK
F	127	372	399	607	89	270	254	437
N	235	331	552	843	194	269	455	707
	každej				***kterej***			
	nekorig.		korig.		nekorig.		korig.	
	BMK	PMK	BMK	PMK	BMK	PMK	BMK	PMK
F	132	363	392	594	132	363	392	594
N	144	135	323	306	144	135	323	306

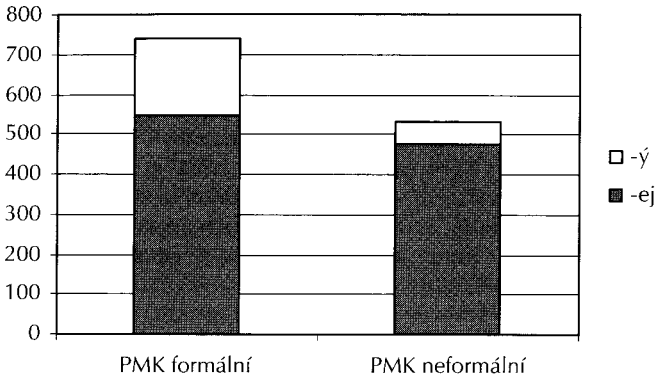
Jak lze konstatovat, korekce značně působí na výsledky, dokonce do takové míry, že může dojít k úplnému převrácení poměru. Tvary *takovej* a *nějakej* mají například nejvyšší absolutní hodnoty v PMK před korekcí v kategorii FORMÁLNÍ a po ní v kategorii NEFORMÁLNÍ. Výsledky při korekci se tak vyrovnávají a nejvyšší absolutní hodnoty jsou na konci všude v kategorii NEFORMÁLNÍ. Toto všechno však platí jenom pro tvary *takovej* a *nějakej*, a kupodivu ne pro *každej* a *kterej*. V tomto případě jsou totiž nejvyšší absolutní hodnoty pořád v kategorii FORMÁLNÍ, před korekcí a také po ní. Nabízí se tedy otázka, zda formálnost působí jinak na užívání koncovek *–ej* podle jednotlivých lexémů. Avšak není tomu tak. Jak se ukáže níže na základě relativních měření a po otestování statistické významnosti, formálnost je nejvýznamnějším rysem a jeho působení na užívání koncovek *-ej* je vždy pozitivní pro kategorii FORMÁLNÍ. Jak tedy vysvětlit, že existuje pro některé lexémy tak značný rozdíl mezi hodnotami absolutními a relativními? Důvodem toho je, že se některá adjektiva vyskytují častěji v určitém typu projevu. Tuto zvláštnost lze ukázat specifickým vyhledáním. V následující tabulce jsou označeny korigované a nekorigované hodnoty v PMK u adjektiv *takový, nějaký, každý, celý, který, dobrý* a *jiný* pro kategorie FORMÁLNÍ a NEFORMÁLNÍ. Tyto hodnoty se netýkají jenom koncovek *-ej*, ale všech adjektivních tvarů. V posledním sloupci je spočítán poměr F/N mezi korigovanými hodnotami v kategoriích FORMÁLNÍ a NEFORMÁLNÍ:

Dotaz v PMK	**nekorigované hodnoty**		**korigované hodnoty**		
	F	**N**	**F**	**N**	**F/N**
takov.*	2531	1495	2097	1885	1,1
ňák.* | něak.* | nějak.*	2083	1178	1726	1485	1,2
každ.*	803	262	665	330	2,0
cel.*	351	214	291	270	1,1
kter.*, ker.*	2614	600	2166	756	2,9
dobr.*	523	343	433	432	1,0
jin.*	764	296	633	373	1,7

Jak lze vidět, rozdíl mezi užíváním v projevech F a N většinou není podstatný (poměr F/N = 1), právě kromě adjektiv *každý* a *který*, pro která je počet tvarů dvakrát anebo třikrát větší v kategorii FORMÁLNÍ než v kategorii NEFORMÁLNÍ (poměr F/N = 2 anebo 3). Největší rozdíl je u vztažného zájmena *který*, což se dá vysvětlit asi tím, že v neformálních projevech je větší konkurence se substandardním univerzálním zájmenem *co*. V následující tabulce jsou označeny absolutní hodnoty pro vztažné zájmeno *který* spolu s procentem užívání koncovky *-ej* v kategorii FORMÁLNÍ a NEFORMÁLNÍ pro PMK:

který	**%**	***-ej***	***-ý***	**celkově**
PMK formální	73,8	547	194	741
PMK neformální	89,7	478	55	533

Převrácení žebříčku mezi absolutními hodnotami a výsledky v procentech se dá vysvětlit tím, že celkový počet tvarů není v kategorii FORMÁLNÍ stejný jako v kategorii NEFORMÁLNÍ. Přestože procento je menší, výsledek je obrácený, protože se aplikuje na větší hodnotu: 73,8 % ze 741 (= 547) je větší než 89,7 % z 533 (= 478), což lze vidět jasně v následujícím grafu:



Uvedený příklad znázorňuje rozdíly týkající se lexémů a parametru formálnosti. Podobné jevy se mohou vyskytovat i na jiných úrovních, kupříkladu u gramatických kategorií, a pro jiné parametry než formálnost (viz o tom Šonková 1999, 194–195). Rozdíly nejsou ovšem všude stejně velké jako ve výše uvedeném příkladu a nemusí vždy dojít k úplnému převrácení mezi hodnotami absolutními a relativními. Potenciálně se však může vždy vyskytovat taková víceméně skrytá nerovnováha, která působí nepřijatelné distorze, a bylo by proto rozumnější se absolutnímu měření vyhnout.

3.3. Měření relativní

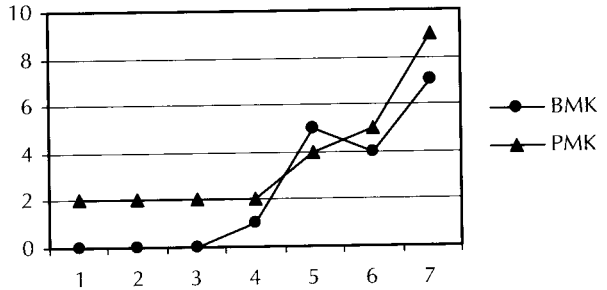
Výhody relativního měření ve srovnání s absolutním jsou jasné, jelikož všechny nedostatky, o kterých právě byla řeč, se ho netýkají. Jeho nevýhoda je spíš rázu čistě praktického a vyplývá z toho, že je potřeba provést druhou řadu měření. V našem případě je tato druhá řada dokonce mnohem pracnější. Zatímco se údaje získané v korpusu o tvarech končících na *-ej* dají zkontrolovat docela lehce, protože stačí škrtnout z frekvenčního seznamu nežádoucí tvary jako *dej, olej* apod., je to v případě tvarů na *-ý* složitější, protože se kvůli silné homonymii koncovek *-ý* musí všechny věty ověřit jedna po druhé.

Výsledky relativního měření mají význam bez ohledu na velikost korpusu a lze je za určitých podmínek (viz výše) přímo srovnávat s jinými výsledky stejného druhu. Toto však neznamená, že výsledky na velikosti korpusu vůbec nezávisí. Velikost korpusu je totiž velmi důležitým faktorem co se týče testování statistické významnosti, jak o tom bude řeč níže (7.).

3.4. Individuální počítání vs globální

Globální rozbor má tu nevýhodu, že může eventuálně skrývat individuální rozdíly. V případě koncovek *-ej* takové rozdíly sice existují, ale nejsou podstatné. Nejmarkantnější je případ osobního zájmena *který*, o němž jsme se již zmínili výše. U jiných jevů to může být samozřejmě jiné, například pro diftongizaci na *-ej* uvnitř slov je možné zaznamenat mezi lexémy daleko větší rozdíly (Krčmová

1997, 225). Ideální metodou by tedy bylo provést individuální sčítání pro všechny lexémy nacházející se v korpusu. Tento systém se však dá systematicky aplikovat jedině pro absolutní měření, a to z praktických důvodů: relativní metoda je totiž omezena na lexémy s vyšší frekvencí, které disponují dostatečným počtem tvarů na *-ej* a zároveň odpovídajících tvarů na *-ý*. Tato omezenost je v našem případě ještě silnější z toho důvodu, že je potřeba systematicky korigovat všechny hodnoty kvůli nerovnováze ve stratifikaci korpusů a provést, jak bylo vysvětleno výše, zvláštní rozbor v každé subkategorii zvlášť. Kvůli tomu je individuální rozbor přijatelný pro velmi omezený počet lexémů. V následující tabulce je označen počet nulových výskytů odpovídajících tvarům na *-ý* pro nejběžnější adjektiva na *-ej* nacházející se v BMK a v PMK:



Jak lze vidět, počet nulových výskytů po prvních třech lexémech velmi rychle stoupá. V této studii jsme arbitrárně omezili rozbor na lexémy, pro které tento počet nepřesahuje v obou korpusech polovinu celkového počtu subkategorií (tj. číslo 8=16:2). Po této selekci se počet lexémů redukoval na pouhých sedm prvků: *takový, ňáký / nějaký / něaký, každý, celý, který / který, dobrý, jiný*. Uvědomíme-li si, že diftongizace na *-ej* je podle všech průzkumů od Kučery (1955) nejběžnějším rysem obecné češtiny, z výše uvedeného sondování jasně vyplývá, že rozsah BMK a PMK není dostačující k tomu, aby se mohly provádět tak jemné rozbory.

Individuální rozbor takto omezený na nejběžnější lexémy má však určité nevýhody, protože se bere v úvahu malý počet příkladů, které mohou vykazovat zvláštní chování. Zdá se totiž, že právě u nejfrekventovanějších slov může docházet k určité lexikalizaci jevů (viz například diftongizace na *–ej* uvnitř slov, o které již byla řeč výše). Na druhé straně omezený rozsah zkoumaných korpusů nedovoluje, jak jsme právě viděli, aby byla provedena individuální analýza lexémů s příliš nízkou frekvencí, která by ideálně umožnila zajímavé porovnání. Nezbude tedy než porovnávat individuální rozbor nejfrekventovanějších slov s rozborem globálním, ve kterém se rozdíly v chování nějakým způsobem kompenzují.

Z hlediska praktického může globální počítání vypadat jednoduše v tom smyslu, že je potřeba provést měření jen jednou. V případě relativního měření je však globální přístup stejně náročný, protože se ověření dat netýká jenom menší skupiny lexémů, ale úplného seznamu výskytů, které lze získat z korpusu. Údaje

v následující tabulce ukazují, že skupina sedmi lexémů, které jsme si k tomu rozboru vybrali, odpovídá v případě koncovek *–ý* v BMK asi třetině výskytů. Tímto sondováním lze odhadnout rozdíl mezi oběma metodami z hlediska pracnosti, ale zároveň i reprezentativnost individuálního sčítání ve srovnání s globálním:

	BMK
takový	1021
ňáký/nějaký/něaký	659
každý	159
celý	98
který/který	476
dobrý	471
jiný	101
celkově	2985
všechna adjektiva na *-ý*	8891

Jak bylo již naznačeno v úvodu, bude provedeno v této studii dvojité měření, nejdřív individuální pro uvedenou skupinu lexémů, a potom globální za účelem porovnání výsledků a možností obou metod z hlediska statistické významnosti.

4. Statistické hodnocení údajů

V předešlých úvahách jsme definovali metodu, jak provádět měření v rámci korpusy nabízených možností. To, že se spokojíme s úrovní spolehlivosti metody však neznamená, že výsledky musí být významné. K problému statistické významnosti uvádíme názornou ukázku vybranou ze soustavy hodnot získaných pro tuto studii. Jedná se o počet výskytů tvarů *takový* a *takovej* v BMK podle parametru věku, který podáváme v následující tabulce po korekci podle výše uvedené metody:

	takovej	*takový*	*% takovej*
IUNIOR	178	54	76,90%
VETUS	173	63	73,20%

Jak lze vidět v posledním sloupci tabulky, tvar *takovej* má v procentech určitou převahu v kategorii IUNIOR. Dá se z toho vyvodit, že věk mluvčího pozitivně ovlivňuje užívání tvaru *takovej*? Rozdíl v procentech ovšem tuto hypotézu podporuje, ale je potřeba se také zeptat, zda tento rozdíl není výsledkem náhody. Na takové otázky lze odpovědět pomocí statistických testů. Vzhledem k tomu, že zde jde o výběr mezi dvěma možnostmi (*-ý* anebo *–ej*), jinak řečeno o tzv. proměnné kategoriální, je vhodné používat testování známé pod jménem χ^2 (chí-kvadrát). Test chí-kvadrát je založen na porovnání mezi hodnotami pozorovanými a očekávanými, kde očekávané jsou četnosti za předpokladu, že proměnné jsou nezávislé. Jsou-li data uspořádaná do tzv. kontingenční tabulky,

	IUNIOR	VETUS	celkový počet
takovej	178 (174)	173 (178)	352
takový	54 (58)	63 (59)	117
celkový počet	232	237	469

můžeme očekávané četnosti pro každé políčko tabulky spočítat podle obecného vztahu (výsledek je označen v závorkách):

očekávaná četnost = (součet v řádku x součet ve sloupci) / celkový počet pozorování
Hodnota testového kritéria χ^2 se potom počítá podle obecného vzorce:
$\chi^2 = \Sigma$[pozorovaná četnost – očekávaná četnost)^2/ očekávaná četnost]

Kritická hodnota testového kritéria χ^2 s hladinou významnosti 0,05 a s jedním stupněm volnosti se rovná 3,84. Vypočtená hodnota v uvedené ukázce $\chi^2 = 0,84$ nepřekračuje kritickou hodnotu, a nelze tedy prohlásit, že na hladině významnosti $\alpha = 0,05$ existuje dostatečný důkaz vztahu mezi věkem a užíváním tvaru *takovej* v Brně. Tím ovšem netvrdíme, že věk nemá vliv, ale pouze konstatujeme skutečnost, že na základě provedených měření si nemůžeme být jistí, že to takhle je.

Testování statistické významnosti je v experimentálních vědách, jako např. v biologii, v medicíně apod., dávno normální praxí. V jazykovědě je to daleko méně obvyklé, snad kromě anglosaské sociolingvistiky, a statistiky se bohužel často používají stále tzv. „naivním" způsobem. V této studii jsme ověřili $\chi 2$ testem všechny údaje získané na ukázku o adj. koncovkách *-ej*.

5. Individuální rozbor lexémů

5.1. Globální hodnoty

První řada měření byla provedena pro sedm vybraných lexémů (*takový, ňáký / nějaký / něaký, každý, celý, který / který, dobrý, jiný*). Následující tabulka uvádí globální hodnoty korigované podle výše uvedeného systému (2):

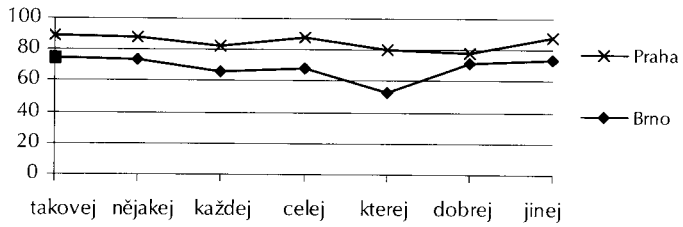
	BMK	PMK		BMK	PMK	χ^2
takovej	352	732	***takový***	117	94	40,37
ňákej/nějakej/něakej	262	578	***ňáký/nějaký/něaký***	95	75	37,72
každej	264	454	***každý***	136	97	33,68
celej	102	170	***celý***	48	23	20,74
kterej/kerej	184	518	***který/který***	162	126	81,05
dobrej	77	141	***dobrý***	30	39	1,49
jinej	72	119	***jiný***	26	16	8,28

Kritická hodnota testového kritéria χ^2 s hladinou významnosti 0,05 a s jedním stupněm volnosti se rovná 3,84. Výsledky označené v tabulce týkající se země-

pisného parametru ukazují, že vypočítané hodnoty všude značně překračují kritickou hodnotu kromě v případě adjektiva *dobrý*. Je tedy možné prohlásit, že původ mluvčího ve většině případů významně ovlivňuje užívaní tvarů na *-ej*. Jak vyplývá z následující tabulky obsahující hodnoty v procentech, rozdíl mezi BMK a PMK je v případě adjektiva *dobrý* menší než u jiných adjektiv, což částečně vysvětluje, proč je hodnota testového kritéria χ^2 menší než kritická hodnota.

	% BMK	% PMK
takovej	75,0	88,6
nějakej	73,4	88,4
každej	66,0	82,3
celej	67,9	88,0
kterej	53,2	80,5
dobrej	71,7	78,5
jinej	73,7	88,0
průměr	68,7	84,9
směrodatná odchylka	7,6	4,3

Jak je znázorněno na následujícím obrázku, přiblížení mezi procenty užívání tvarů na *-ej* v BMK a v PMK pro adjektivum *dobrý* není dáno tím, že procento je vyšší v Brně, ale spíš že je o něco menší v Praze. Jazyková interpretace tohoto jevu není jasná a není vyloučeno, že to může být jen výsledkem náhody. V případě vztažného zájmena *který* lze konstatovat, že rozdíl mezi procenty je naopak větší než u jiných adjektiv, což by se dalo interpretovat tak, že *který* konkuruje v mluvených projevech substandardnímu tvaru *co*.



5.2. Vliv sociolingvistických parametrů

Provedená měření v jednotlivých subkategoriích nám umožňují spočítat korigované hodnoty jednotlivě pro všechny sociolingvistické parametry. Všechny tyto hodnoty byly otestovány testem χ^2 a výsledky jsou uvedeny v následující tabulce:

	věk				pohlaví				formálnost				vzdělání			
	BMK		PMK		BMK		PMK		BMK		PMK		BMK		PMK	
	x2		x2		x2		x2		x2		x2		x2		x2	
takovej	0,84	ne	9,97	ano	4,27	ano	0,48	ne	18,64	ano	22,57	ano	5,46	ano	4,56	ano
nějakej	1,01	ne	5,57	ano	0,61	ne	4,22	ano	15,63	ano	11,90	ano	16,64	ano	3,01	ne
každej	2,56	ne	5,66	ano	1,47	ne	13,87	ano	22,15	ano	4,18	ano	31,00	ano	22,34	ano
celej	4,93	ano	0,53	ne	0,32	ne	0,00	ne	13,55	ano	3,47	ne	1,14	ne	2,35	ne
kterej	7,28	ano	10,61	ano	32,01	ano	4,03	ano	34,39	ano	25,25	ano	29,02	ano	0,84	ne
dobrej	1,51	ne	1,84	ne	5,80	ano	2,99	ne	10,02	ano	1,18	ne	0,40	ne	0,58	ne
jinej	3,59	ne	3,31	ne	4,46	ano	2,41	ne	16,29	ano	10,63	ano	2,26	ne	2,65	ne
P	5		3		3		4		0		2		3		5	

Kritická hodnota, která je jako předtím rovna 3,84, zde není v mnoha případech dosažena (hodnota P v posledním řádku naznačuje počet údajů v každém sloupci, které nejsou statisticky významné). Důvody těchto negativních výsledků jsou následující:

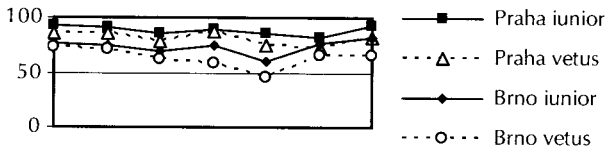
– rozdíly mezi hodnotami jsou relativně malé. Je příznačné, že největší počet pozitivních výsledků se týká parametru formálnosti, který se vyznačuje právě podstatnějšími rozdíly.
– počet výskytů v jednotlivých subkategoriích je relativně omezený a tento faktor přímo ovlivňuje hodnotu χ^2 (jednoduše řečeno: čím jsou pozorované hodnoty menší, tím větší je pravděpodobnost, že výsledky jsou dílem náhody).

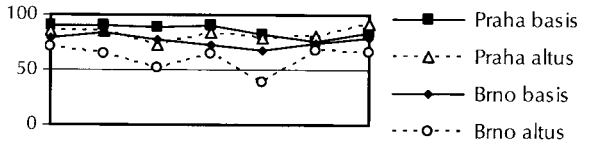
V následující tabulce jsou uvedena procenta užívání tvarů na *-ej* podle jednotlivých sociolingvistických parametrů. Případy, pro které hodnota χ^2 nepřekračuje kritickou hodnotu, jsou označeny šedě:

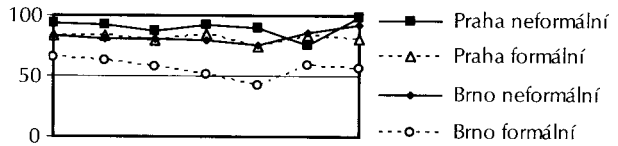
%	věk				pohlaví				formálnost				vzdělání			
	BMK		PMK		BMK		PMK		BMK		PMK		BMK		PMK	
	I	V	I	V	M	Z	M	Z	F	N	F	N	A	B	A	B
takovej	76,9	73,2	92,0	85,0	79,2	71,0	89,3	87,8	66,0	83,3	82,8	93,3	70,2	79,6	86,3	91,0
nějakej	75,4	70,6	90,9	84,9	75,1	71,4	90,5	85,3	62,5	81,3	83,2	92,0	64,6	83,7	86,2	90,6
každej	68,8	61,0	86,3	78,6	69,0	63,3	89,0	76,8	57,4	80,5	80,1	87,2	51,3	77,8	73,0	88,6
celej	75,3	58,2	89,9	86,4	70,4	66,0	88,1	88,0	51,6	80,0	84,1	92,8	63,9	72,1	83,8	91,0
kterej	60,6	46,1	85,6	75,4	67,0	36,5	83,0	76,5	42,0	75,4	73,8	89,7	38,7	67,6	79,2	82,1
dobrej	76,0	65,0	82,0	73,6	79,7	58,0	73,2	83,8	58,5	86,1	82,8	75,9	68,3	73,9	80,6	75,9
jinej	82,0	65,1	92,8	82,6	82,7	63,9	93,0	84,2	55,9	91,9	80,2	98,6	66,4	79,8	92,4	83,3
průměr	73,6	62,7	88,5	80,9	74,7	61,4	86,6	83,2	56,3	82,6	81,0	89,9	60,5	76,4	83,1	86,1
S	6,9	9,0	4,0	5,1	6,1	11,9	6,6	4,7	7,8	5,2	3,5	7,1	11,4	5,5	6,2	5,8

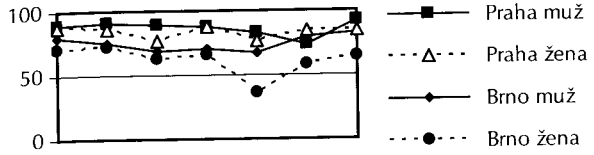
Případy, ve kterých se statistická významnost nedá prokázat, někdy odpovídají minimálnímu rozdílu v procentech, například pro adjektivum *celý* u parametru pohlaví v PMK: 88,1 % u mužů a 88 % u žen. Jinde zase rozdíl činí dokonce až 10 %, jako pro adjektivum *jiný* u parametru věku v PMK: 92,8 % u mladých a 82,6 % u starších. Jedná se tu ovšem o diametrálně odlišné situace, protože hodnota χ^2 je v prvním případě menší než 0,01, což znamená, že pravděpodobnost je minimální, zatímco v druhém případě je naopak velmi blízká kritické hodnotě (3,58 ve srovnání s 3,84) a pravděpodobnost je jen o málo větší (0,058) než standardní hladina 0,05. V takovém případě by se tedy mohlo stejně prohlásit, že dotyčný parametr je statisticky významný. Není bez zajímavosti podívat se taky na prostřední (vyvážené?) situace, například na případ adjektiva *nějaký* u parametru věku v BMK, kde je procentní rozdíl kolem 5 %. V takových případech by se určitě při „naivní" interpretaci tvrdilo, že dotyčný parametr ovlivňuje užívání tvarů na *-ej*. Statistické testování však ukazuje, že pravděpodobnost, že jsou zjištěné rozdíly pouze výsledkem náhody, se rovná 0,3, což je skutečně příliš vysoká hodnota (jedna možnost ze tří). Chyba „naivní" interpretace tkví v tom, že se berou v úvahu procenta nehledě na velikost pozorovaných hodnot, zatímco tento faktor je při testování rozhodující.

V následujících grafech lze najít grafické znázornění výše uvedených výsledků v procentech:









6. Celokorpusový rozbor

6.1. Globální hodnoty

Výskyty se zde počítají globálně pro všechna adjektiva v korpusu, a získané hodnoty jsou proto mnohem vyšší než v případě individuálního rozboru podle jednotlivých lexémů. Výsledkem toho je, že vypočítaná hodnota χ^2 značně překračuje kritickou hodnotu 3,84, a lze prohlásit, že zeměpisný faktor má globálně statisticky významný vliv na užívání adjektivních tvarů na *-ej*:

	BMK	PMK		BMK	PMK	χ^2
adjektiva na *-ej*	2845	6567	**adjektiva na *-ý***	1433	1132	577,44

V následující tabulce jsou označena procenta spočítaná podle korigovaných a nekorigovaných hodnot:

%	BMK	PMK
nekorigované hodnoty	70,0	84,6
korigované hodnoty	66,5	85,3

O rozdílech mezi hodnotami korigovanými a nekorigovanými lze říci, že jsou v souladu s tím, co bylo výše poznamenáno o nerovnováze ve stratifikaci korpusů. Brněnský korpus se například jevil jako příliš neformální, a je tedy logické, že korekcí dochází ke zmenšení procenta tvarů na *-ej*. Pražský korpus obsahuje naopak příliš moc formálních textů a korekce přirozeně působí opačným směrem. To, že vliv korekce je podstatnější pro brněnský korpus se dá vysvětlit částečně tím, že nerovnováha stratifikace je v něm výraznější, ale také větším vlivem sociolingvistických parametrů v Brně, hlavně formálnosti. O tom všem bude řeč podrobně v následujícím odstavci.

6.2. Vliv sociolingvistických parametrů

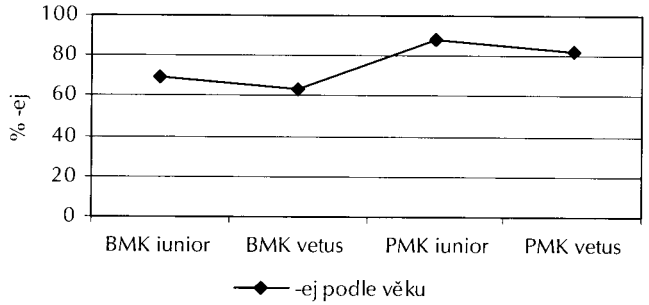
Vypočítané hodnoty χ^2 pro jednotlivé sociolingvistické parametry jsou v celokorpusovém rozboru vyšší než při rozboru podle jednotlivých lexémů. Ve všech možných kombinacích tyto hodnoty značně překračují kritickou hodnotu 3,84, což prokazuje, že rozdíl v užívání tvarů na *-ej* je na úrovni celého korpusu statisticky významný pro všechny analyzované parametry.

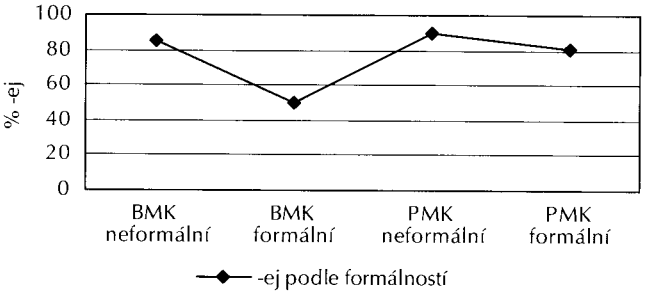
věk				pohlaví				formálnost				vzdělání			
BMK		PMK		BMK		PMK		BMK		PMK		BMK		PMK	
χ^2		χ^2		χ^2		χ^2		χ^2		χ^2		χ^2		χ^2	
46,99	ano	102,83	ano	144,09	ano	31,46	ano	1598,89	ano	280,91	ano	302,41	ano	137,71	ano

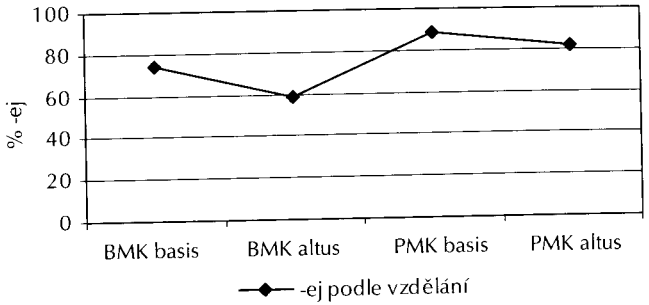
Hodnoty vyjádřené v procentech ukazují, že tento vliv působí kladně na užívání tvarů na *-ej* u mládeže, u mužů, v neformálních situacích a u osob s nižším vzděláním, a to stejně v BMK jako v PMK:

věk				pohlaví				formálnost				vzdělání			
BMK		PMK		BMK		PMK		BMK		PMK		BMK		PMK	
I	V	I	V	M	Z	M	Z	F	N	F	N	A	B	A	B
69,2	63,2	88,2	82,4	71,6	61,1	86,7	83,5	50,2	85,4	80,6	90,2	58,8	74,0	81,9	88,7

Grafické znázornění lze najít v následujících schématech:







6.3. Interpretace

6.3.1. Věk

Na vliv věku v Praze upozornila již J. Šonková ve své studii o mluvené češtině, avšak bez statistických údajů (1999, 192), a před ní se o tomto faktoru všeobecně vyjadřovali Sgall et al. (1992, 23–24). Náš průzkum ukazuje, že vliv věku na užívání tvarů na *-ej* je v Praze a v Brně přibližně stejný a činí zhruba 4 %. Tento výsledek si zaslouží několik poznámek:

– Minimální věk mluvčích, jejichž projevy jsou vloženy do korpusu, je 20 let. Tím se vylučují eventuální interference s parametrem vzdělání v tom smyslu, že všechny dotyčné osoby ukončily povinnou školní docházku, ale zůstává úplně stranou velmi zajímavá doba dospívání.
– Vliv věku nelze přímo interpretovat s ohledem na vývoj systému. Mladí a starší lidé nemluví stejným způsobem z různých důvodů. Velmi pravděpodobně je starší generace konzervativnější než mladší a nepřijímá tak lehce eventuální změny, ale není vůbec řečeno, že je musí systematicky odmítat. Chceme-li pozorovat vývoj, je potřeba provést průzkum pro stejnou věkovou skupinu v různých časových intervalech, např. mládež v roce 1990, 2000, 2010 atd.
– Bylo by potřeba zjistit, jaké je přesné rozložení různých věkových skupin v BMK a v PMK, abychom zkontrolovali, zda každá skupina je dostatečně reprezentována. Větší přítomnost mluvčích středního věku (hranice je okolo 35 let) by totiž mohla setřít rozdíly mezi kategoriemi IUNIOR a VETUS. Ideální by ovšem bylo zkoumat situaci podle různých věkových skupin, což systém označkování v BMK a v PMK neumožňuje a což může být stejně problematické pro nedostačující počet výskytů (viz výše rozbor podle jednotlivých lexémů).
– Ve specifické situaci češtiny je důležitým faktorem i zaměstnání, jak na to správně upozornili Sgall et al. (1992, 22–23, 194–195). Intuitivně se tedy může zdát, že vliv věku je spíš odrazem větší profesionální zkušenosti. Musíme si však být vědomi toho, že není vyloučena opačná situace, když

někdo vůbec nepoužívá spisovný jazyk ve své profesi a pomalu ztrácí schopnost vyjádřit se spisovně, kterou předtím získal ve škole. Určitě by bylo proto zajímavé tento faktor podrobně zkoumat na základě nových korpusů, které by braly v úvahu nejen parametr věku (a vzdělání), ale zároveň i profese.

Na první pohled by se dalo asi intuitivně říci, že vliv věku je příznakem známé tendence mladých lidí vyjadřovat se nekonvenčně. Tento faktor hraje bezesporu určitou roli, ale z předešlých úvah vyplývá, že se tu kombinují i jiné faktory a že takové závěry jsou bez dalších specifikací zatím příliš povrchní.

6.3.2. Pohlaví

Průzkum ukazuje, že statisticky významný vliv na užívání tvarů na *–ej* má i pohlaví. Jak v Brně, tak v Praze se tento vliv projevuje tím, že substandardní tvary jsou běžnější u mužů, avšak s tím rozdílem, že tato tendence je silnější na Moravě než v Čechách (10 % v Brně a ani ne 3 % v Praze). J. Šonková věnovala těmto genderovým rozdílům průkopnický článek (Šonková 1999), ale zabývala se spíš všeobecně mluvnickými kategoriemi a slovní zásobou a nechala trochu stranou specifickou problematiku obecné češtiny. Tato problematika se stala nedávno středem pozornosti Z. Hladké (2001, 2004) na základě analýzy velmi zajímavého korpusu soukromé korespondence. Autorka ukázala mimo jiné, že u žen existuje určitá tendence vzdálit se od spisovné normy ve specifickém případě soukromé korespondence. Na základě sondování v BMK zároveň upozorňovala na to, že je to v mluvené řeči spíš naopak (Hladká et al. 2004). Toto obrácení se podle Z. Hladké vysvětluje tím, že ženy se vyznačují větší spontánností v soukromé korespondenci , a proto častěji využívají rysy, které jsou typické pro běžný rozhovor.

Tendence více využívat substandardních tvarů na *-ej* u mužů v mluvených projevech, pozorovaná již Z. Hladkou, byla zde ověřena statistickým testováním a přesněji měřena jak v brněnské, tak v pražské jazykové situaci. Všeobecně lze říci, že takové poznatky jsou možná novinkou v českém prostředí (gender není citován mezi tzv. individuálními faktory u Sgalla et al. 1992), ale nejsou tak překvapující ve světle existující odborné literatury o jiných sociolingvistických situacích. Zdá se totiž, že existuje docela rozšířená tendence, podle níž muži mají ve stabilizovaných sociolingvistických situacích větší sklon používat substandardní tvary (Labov 1990, Cheschire 2002).

6.3.3. Vzdělání

Úroveň vzdělání má statisticky významný vliv, jak v Brně (v relativně velké míře, zhruba 15 %), tak v Praze (7%). Jak bylo řečeno výše, PMK a PMK berou v úvahu jenom mluvčí starší než 20 let. Předešlé úvahy o vlivu věku platí i zde v tom smyslu, že by bylo zajímavé prozkoumat, do jaké míry má vysokoškolské vzdělání vliv na pozorované jevy ve srovnání se zaměstnáním. Číselné údaje o vlivu vzdělání jsou totiž příliš všeobecné a bylo by potřeba

provést různé konfrontace s jinými parametry. Tento přístup bude znázorněn níže v konfrontačním rozboru, který bere v úvahu různé kombinace parametrů uvnitř kategorie formálnosti, a bude mimochodem vidět, že přinese cenné informace právě o roli vzdělání.

6.3.4. Formálnost

Formálnost projevu je podle průzkumu nejvýznamnější parametr. Procentuální rozdíl mezi formálními a neformálními projevy je však mnohem větší v Brně než v Praze (35 % oproti 10 %). V Praze je vliv formálnosti jen o málo větší než u jiných parametrů, jelikož užívání tvarů na *-ej* zůstává i ve formálních situacích velmi časté (nad 80 %). V případě Brna jsou velmi pozoruhodné dva fakty:

(i) užívání tvarů na *-ej* pří neformálních projevech je v Brně skoro na stejné úrovni jako v Praze (85 % oproti 90 %);
(ii) tento vliv velmi prudce klesá (pod 50 %) ve formálních situacích.

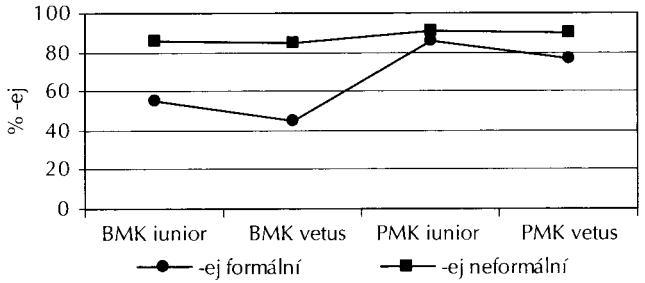
Tento jev je obzvlášť zajímavý a zaslouží si větší pozornost. Jednotlivá měření ve všech subkategoriích nám dovolují zachytit korigované hodnoty pro různé kombinace parametrů a zjistit, jestli jsou v BMK a v PMK významné rozdíly uvnitř opozice FORMÁLNÍ – NEFORMÁLNÍ pro jiné parametry, jako např. věk, vzdělání anebo pohlaví. Vypočítané hodnoty χ^2 jsou pro tyto kombinace následující:

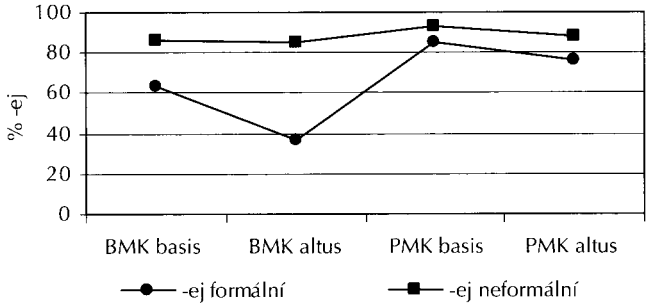
formálnost +	BMK		PMK	
	χ^2		χ^2	
IUNIOR	273,73	ano	32,08	ano
VETUS	321,57	ano	112,02	ano
BASIS	140,60	ano	62,48	ano
ALTUS	499,45	ano	83,77	ano
MUŽ	179,42	ano	52,52	ano
ŽENA	437,29	ano	93,26	ano

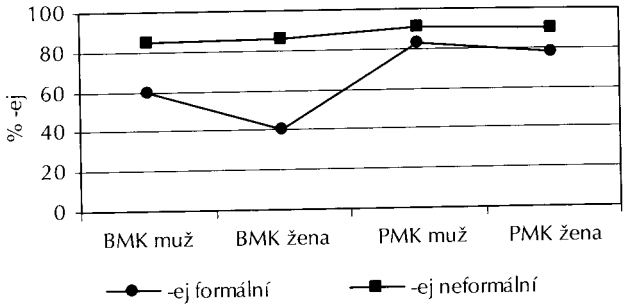
Jak lze konstatovat, všechny hodnoty χ^2 jsou vyšší než kritická hodnota 3,84, a lze tedy prohlásit, že rozdíly jsou statisticky významné pro všechny rozebrané kombinace parametrů. Tyto rozdíly jsou vyznačeny v procentech v následující tabulce:

	% BMK	% PMK
IUNIOR + FORM	54,6	85,2
VETUS + FORM	44,7	76,2
IUNIOR + NEFORM	86,2	91,1
VETUS + NEFORM	84,3	89,3
ALTUS + FORM	36,7	76,3
BASIS + FORM	63,6	84,8
ALTUS + NEFORM	84,8	87,7
BASIS + NEFORM	85,9	92,9
MUŽ + FORM	59,5	83,1
ŽENA + FORM	40,6	77,5
MUŽ + NEFORM	85,3	90,6
ŽENA + NEFORM	85,4	89,8

Grafické znázornění těchto rozdílů ukazuje jasněji, jaký mají vliv jednotlivé parametry v obou městech:







Co se týče Prahy, lze konstatovat, že rozdíly jsou relativně drobné. Je poznat pouze slabý pokles rozdílu uvnitř kategorie FORMÁLNÍ – NEFORMÁLNÍ týkající se mladých (minus 8 %), mužů (minus 3 %) a nevzdělaných lidí (minus 9 %). V Brně se rýsuje podobný pokles, jenže výraznější a rozrůzněný podle jednotlivých parametrů. V případě věku je totiž procentní rozdíl skoro stejný jako v Praze (minus 7 %), pak ale značně stoupá u mužů (minus 19 %), a hlavně u nevzdělaných lidí (minus 26 %). Výsledky průzkumu se dají shrnout takto:

– rozdíl užívání tvarů na *-ej* podle formálnosti je v Praze relativně drobný a málo závisí na jiných faktorech.
– rozdíl užívání tvarů na *-ej* podle formálnosti je v Brně podstatnější a jasně spjatý s kritériem vzdělání, v menší míře i pohlaví (viz taky Hladká-Šindlerová 2004, 110).

6.3.5. Závěr

Závěrem globálního rozboru podle sociolingvistických parametrů lze říci, že tvary na *–ej* se výrazně uplatňují na Moravě, dokonce do takové míry, že je brněnští mluvčí užívají v neformálních situacích skoro stejně často jako v Praze. Pokud je však situace formální, stanou se tyto tvary pro většinu mluvčích nepřijatelnými, hlavně u vzdělaných lidí a u žen, které jsou tu značně konzervativnější. Vliv věku je naopak slabší a přibližně na stejné úrovni jako v Praze. Korpusový rozbor tedy podává poněkud složitý obraz. Pronikání tvarů obecné češtiny na Moravu je skutečné a kvantitativně velmi pozoruhodné, ale neimplikuje, jak by to mohlo vypadat na první pohled, že se sociolingvistické situace v Brně a v Praze přibližují. Tvary na *-ej* jsou sice přítomné v obou městech, ale mají docela odlišnou platnost: zatímco v Praze jsou rysem **mluvenosti** (používají se skoro ve všech situacích a formálnost na to nemá velký vliv), v Brně jsou hlavně rysem **neformálnosti** (vzdělaní lidé se jim ve formálních situacích vyhýbají). Tyto rozdíly by mohly ovšem časem vymizet, ale to je zatím pouhá hypotéza.

7. Perspektivy

Prvním cílem této studie bylo analyzovat meze a možnosti srovnávacího rozboru mluvených korpusů na základě statistických údajů. Po úvahách o povaze těchto korpusů, o typu měření a o významnosti získaných údajů vyšlo najevo několik překážek k uskutečnění takového rozboru:

- stratifikace korpusu by měla být přesná i z kvantitativního hlediska, aby se nemusela provádět systematická korekce údajů;
- je potřeba vyhnout se absolutnímu měření vzhledem k potenciálním distorzím na různých úrovních;
- je nutné ověřit statistickou významnost údajů specifickými testy.
- je těžké získat statistické významné výsledky v případě jemnějších rozborů, např. na úrovní lexémů, kvůli relativně omezenému rozsahu analyzovaných korpusů.

Snažili jsme se zjistit na základě příkladu adjektivní koncovky *-ej*, jaké důsledky má nerespektování těchto pravidel. Provedená číselná hodnocení ukazují, že nejsou vůbec zanedbatelné. Bohužel v korpusových studiích se příliš často tato základní pravidla ignorují a provádí se bez jakékoliv korekce srovnávání mezi údaji pocházejícími z korpusů, které mají úplně jinou strukturu. Ještě častěji se budují teorie na základě velmi omezených počtů výskytů čerpaných z malých vzorků, které velmi pravděpodobně nemají žádnou statistickou významnost.

Co se týče konkrétního příkladu, který jsme zde analyzovali jako názornou ukázku, získané výsledky svědčí o tom, že kvantitativní rozbor stratifikovaných korpusů je sice pracný a složitý, ale přinese velmi cenné poznatky o sociolingvistické situaci. Dokázali jsme například, že pronikání obecné češtiny je v případě adjektivní koncovky *-ej* velmi výrazné, ale že dotyčné tvary mají v Brně zatím jinou platnost než v Praze, protože jsou na Moravě pořád velmi silným příznakem neformálnosti. Mnoho dalších statistických údajů se však dá interpretovat jenom částečně a vyšlo najevo, že rozsah korpusů tu není dostatečný nebo že by bylo potřeba zahrnout do struktury další parametry, jako například druh zaměstnání.

Literatura

Berruto G., 1987, *Sociolinguistica dell'italiano contemporaneo*. La Nuova Italia Scientifica, Roma.

Butler C. 1985, *Statistics in Linguistics*. Blackwell, Oxford.

Cheshire J., 2002, Sex and gender in variationist research. In J. K. Chambers, P. Trudgill and N. Schilling-Estes (eds.), *Handbook of Language Variation and Change*. Blackwell, Oxford, 423–43.

Čermák F., 2006, *Mluvené korpusy*, zde 53–67

Davis L.M., 1990, *Statistics in Dialectology*. University of Alabama Press, Tuscaloosa.

Gadet F., 2003, *La variation sociale en français*. Ophris, Paris.

Hammer, L..: Code-switching in Colloquial Czech, In J. L. Mey (ed.), *Language and Discourse: Test and Protest*. Benjamins, Amsterdam, 455–473.

Hladká Z., 2001, Spisovnost a nespisovnost v jazyce soukromé korespondence (se zřetelem k teritoriální příslušnosti pisatelů). *Naše řeč*, 84, 225–234.

Hladká Z, H. Šindlerová, 2004, Jakou češtinou si dopisujeme na Moravě. In *Acta Universitatis Palackianae Olomucensis, Fac.Phil., Moravica 1*. Univerzita Palackého v Olomouci, Olomouc, 105–113.

Hladká Z., 2004, Soukromá korespondence z hlediska rodových diferencí. In: *Súčasná jazyková komunikácia v interdisciplinárnych súvislostiach*. Univerzita Mateja Bela, Banská Bystrica, 469–475.

Hladká Z. 2005, Zkušenosti s tvorbou korpusů češtiny v UJČ FF MU v Brně. In *Sborník prací Filozofické fakulty brněnské univerzity A 53*, Masarykova univerzita, Brno, 115–124.

Kopřivová M., 2004, Srovnání užívání některých adjektivních koncovek v psaném a mluveném jazyce (na materiálu Českého národního korpusu). In: *Spisovnost a nespisovnost. Zdroje, proměny a perspektivy*. Pedagogická fakulta MU, Brno, 167–171.

Krčmová M., 1981, *Běžně mluvený jazyk v Brně*. Univerzita J. E. Purkyně, Brno.

Krčmová M., 1997, Současná běžná mluva v českých zemích. In F. Daneš et al.: *Český jazyk na přelomu tisíciletí*. Academia, Praha, 160–172.

Krčmová M., 1997, Proměny brněnské městské mluvy, In F. Daneš et al.: *Český jazyk na přelomu tisíciletí*. Academia, Praha, 225–230.

Kretzschmar W. A., E. W. Schneider, 1996, *Introduction to Quantitative Analysis of Linguistic Survey Data*, Sage Publications, Thousand Oaks – London – New Delhi.

Kravščinová K., B. Bednářová,1968, Z výzkumu běžné mluvené češtiny. *Slavica Pragensia*, 10, 305–320.

Labov W., 1990, The intersection of sex and social class in the course of linguistic change. *Language Variation and Change* 2, 205–254.

Maglione C., 2003, Remarks on new research in everyday Czech. *Prague Bulletin of Mathematical Linguistics* 79–80, 87–100.

Oakes M. P., 1998, *Statistics for Corpus Linguistics*. Edinburgh University Press, Edinburgh.

Sgall P. et al, 1992, *Variation in Language. Code switching in Czech as a challenge for sociolinguistics*. John Benjamins, Amsterdam-Philadephia.

Sgall, P., 2004, K vývoji výzkumu obecné češtiny (OČ). In *Spisovnost a nespisovnost. Zdroje, proměny a perspektivy*. Pedagogická fakulta MU, Brno, 34–39.

Šonková J., 2000, *Mluvená čeština a korpusová čeština, SaS*, 61, 190–202.

Šonková J., 1999, Gender-based Results of a Quantitative Analysis of Spoken Czech: Contribution to the Czech National Corpus. In M. Mills (ed.), *Gender Linguistics*, John Benjamins Publishing Company, Amsterdam – Philadelphia, 183–200.