

Pilotní studie o vztahu mezi korpusovými daty a soudy přijatelnosti konkurujících si tvarů
Neil Bermel
n.bermel@sheffield.ac.uk

Abstract:

A pilot study on the relationship between corpus data and acceptability judgments of competing forms

This article reports on the results of a pilot study in which data on morphological variation in genitive case forms found in the Czech National Corpus was compared with that gathered from a small sample of native speakers of Czech (N=20). The results show that, while there are statistically significant similarities between the two groups of data, the similarities do not allow us to map directly from corpus frequency of the competing forms onto their acceptability or vice versa.

1. Úvod

Současná studie vznikla jako příspěvek k odpovědi na otázku: můžeme interpretovat data z korpusu jako věrnou reprezentaci přijatelnosti zkoumaných jevů pro rodilé mluvčí?¹

Otázka vzniká alespoň částečně proto, že v rámci výzkumu české morfologie jsme používali Český národní korpus, ze kterého dvě rozsáhlé složky jsou vyhlášeny jako „reprezentativní“ korpusy synchronní psané češtiny: SYN2000 a SYN2005. Tyto korpusy byly sestaveny na základě průzkumu o zvycích čtenářů provedeného v 90. letech 20. století a v letech 2000-2001 (viz např. Šulc (2001), Králík (2001)). Korpus takto sestavený má podávat věrný obraz psaného světa, který obklopuje čtenáře. Je možné, dokonce pravděpodobné, že uživatelé jazyka jsou do určité míry ovlivněni tímto obrazem, že vytváří nebo modifikuje jejich „estetické“ reakce a soudy o použitelnosti jazykových jevů. Podobný názor je (často nevyřčeným) předpokladem mnohé korpusové studie.

Tím se nabízí první hypotéza:

Hypotéza 1 („silná“): Frekvence v korpusu odráží nebo vytváří preference rodilých mluvčích pro ten či onen tvar nebo strukturu, a proto je procentuální zastoupení tvarů v korpusu proporcionálně shodné s preferencemi rodilých mluvčích.

Základem této hypotézy je nejčastěji introspekce jednotlivých výzkumníků, tj. jejich korpusová data jsou spolehlivá, ale přijatelnost zkoumaného jevu se opírá o intuici jednoho badatele, spíše než o sondy do názorů více uživatelů.²

Dosud bylo provedeno pouze několik analýz většího rozsahu, ale zatím žádná z nich nepodporuje tuto silnou hypotézu. Halliday (1991a, 1991b, 1992) navrhuje frekvenční poměr 9:1 jako významný z informačního a strukturního hlediska a nabízí proto frekvenční hranice 10 % a 90 % jako obzvlášť významné. Výzkum Kempena a Harbuschové (2005) a Divjakové (2008) poukazuje na asymetričnosti korespondencí mezi korpusovými daty a sondami do názorů rodilých mluvčích v oblasti syntaxe. Podle nich se nabízí méně kategorická verze této hypotézy:

Hypotéza 2 („slabá“): Frekvence v reprezentativním korpusu a preference v něčem souvisejí, ale tuto souvislost nelze přeceňovat. Vztah mezi procentuálním zastoupením v korpusu a preferencemi rodilých mluvčích je statisticky významný; nicméně korelace mezi frekvencemi a preferencemi nejsou striktní a nezdají se být platné ve všech případech.

¹ Výzkum vznikl v rámci projektu „Kapitoly z české gramatiky“ vedeného F. Štíchy.

² Netvrdím, že korespondence „korpusová data – reflexe jednoho lingvisty“ nesvědčí o ničem v systému jazyka, metodologicky jsou ale opodstatněné výhrady proti tomuto přístupu. Lingvista, byť sebezkušenější, sám ví, co našel v korpusu, má podrobnou znalost oblasti, kterou zkoumá, a proto nemůže samotný sloužit jako spolehlivý zdroj informací. Jde zároveň o paradox pozorovatele a známý problém dat získaných od lingvistů.

Je ovšem možná i třetí varianta, ve které korpusové údaje nepřinášejí žádné poznatky o stavu jazyka jako takového:

Hypotéza 3 („nulová“): Frekvence v korpusu nesouvisí s preferencemi rodilých mluvčích. Neexistují žádné významné korelace mezi korpusovou frekvencí jevů a jejich přijatelností pro rodilé mluvčí.

Třetí hypotéza se na první pohled zdá být extrémní, ale lze najít opodstatněné argumenty v prospěch podobných formulací, viz např. Olivu a Doležalovou (2004).

2. Rysy problematiky

Rozhodl jsem se empirickým způsobem pokusit o sondu do vztahu mezi korpusovými daty a přijatelností morfologických jevů. Zajímavý příklad variace se nabízel uvnitř vzoru *hrad*, který se hodil hned z několika důvodů.

Hrad je vysoce frekventovaný vzor se značným rozkolísáním v jednotlivých tvarech. Pro nastínění rozsahu deklinačního typu jsem se opíral o tvary druhého pádu jednotného čísla (dále Gsg), který jasně ukazuje rozdíly mezi deklinacemi (*hrad-u* oproti *stroj-e*). Podle korpusu SYN2005 je evidováno v Gsg celkem 5,5mil. textových slov, která jsou označena jako podst. jména neživ. rodu. Z nich 1,5mil. textových slov patří do vzoru *hrad*, tedy cca 27 %. Podobná je frekvence lemmat (lexémů): zhruba 26tis. lemmat (18 %) patří do vzoru *hrad* z celkově 142tis. lemmat podst. jmen neživ. rodu.

Variace uvnitř paradigmatu *hrad* byla předmětem několika studií od 50. let minulého století. Některé se jenom pokusily o nastínění problematiky na základě preskriptivních a deskriptivních příruček (Cummins 1995, Rusínová 1992, Sedláček 1982) nebo excerpce z knižní produkce (Klimeš 1953); jiné se opíraly o sondy do úzu rodilých mluvčích (Kasal 1992) či o jejich hodnocení variantních tvarů (Bermel 1993). Jiné, pozdější studie probírají rozsah variace v datech čerpaných z Českého národního korpusu (Bermel 2004, Štícha v tisku). Žádná studie se ale zatím nezaměřila na možné souvislosti mezi korpusovými daty a sondami do soudů rodilých mluvčích.

Z různých variačních možností v paradigmatu *hrad* jsem zvolil Gsg, v němž podstatná jména končí na *-u*, nebo *-a*, případně jsou možné oba tyto formanty. Výhodné pro daný účel je, že se v posledním případě volba formantů nejčastěji nezdá být motivována.³ Lexikální podmiňovací faktory jsou velmi slabé: převzatá slova jen málokdy mívají *-a* a homonymie se slovem živ. rodu může ovlivnit volbu. Žádné fonologické podmiňovací faktory neexistují.

O podmiňovacích faktorech syntaktických se dočteme v *Příruční mluvnici češtiny*:

„Substantiva s adverbialním významem místa (méně často i času nebo způsobu) se pojívají s koncovkou *-a*. Jako objektová (nebo v jiných významech) mívají též substantiva koncovku *-u*: *do kalicha/u – dotkl se kalichu, do roka – dožil se jednoho roku, do kouta – nebylo jediného koutu, do rybníka – nebylo jednoho rybníku*“ (PMČ 1995:253).

Příklady citované výše se týkají dvou syntaktických konstrukcí, které jsou v současné češtině poměrně řídké: přímý objekt v 2. pádě a 2. pád při záporu. Pro ověření tohoto tvrzení je proto nutné, aby analýza zahrnovala dostatečný počet rozdílných syntaktických kontextů, včetně záporného genitivu a vazby sloveso + genitiv.

Na začátku bylo účelné zjistit rozsah variace v Gsg.

³ MČ tvrdí, že „Distribuce podob závisí na různých činitelích – slovotvorných, syntaktických, sémantických, popř. i na hláskovém skladu slova“ (1986: II, 305). Mluví se tu však o obecné distribuci tvarů, nikoli o distribuci tvarů u lexémů, kde dochází k variaci.

Tabulka 1. Rozsah variace v Gsg

| | celkem | -u | % u | -a | % a |
|--|---------------|-----------|------------|-----------|------------|
| Textová slova vzoru <i>hrad</i> v SYN2005 ⁴ | 1 374 025 | 1 193 709 | 86,9 | 180,316 | 13,1 |
| Lexémy v SYN2005 | 13 688 | 13 017 | 95,1 | 804 | 5,8 |

Lexémy s formantem *-a* v Gsg tvoří jenom 5,8 % všech lexémů, ale představují 13,1 % všech výskytů Gsg. Obecná frekvence tvarů na *-a* je tedy vyšší, než u tvarů na *-u* (hlavně asi tím, že se formant *-a* neuplatňuje u tolika velmi málo frekventovaných slov).

3. Parametry dotazníku

Při sestavování pilotní studie jsem se zaměřil na lexémy s variací *-u/-a* v 2. pádě j. č. doloženou v SYN2005. Takovou variaci najdeme celkem u 112 podst. jmen. Některá mají vysokou frekvenci (1000+ výskytů), např. *rok, zákon, zákoník, jazyk, kostel, sen*, v této skupině jsou ale i četná podst. jména s nízkou frekvencí (< 10 výskytů). Frekvence formantu *-u* oproti *-a* se značně liší uvnitř této skupiny (tj. nejde o stálou proporcí tvarů s jedním či druhým formantem), ačkoli si všimáme nízkého počtu slov s rovnocennou distribucí formantů.

Vybral jsem celkem sedm slov, která měla představovat různé frekvence výskytů formantů *-a* a *-u* v korpusu. Na krajních pozicích této škály (*les, průchod*) jsou tzv. „kontrolní“ slova, u kterých se v korpusu vyskytl jenom jeden formant Gsg, tedy *lesa*, nikoli **lesu* a *průchodu*, nikoli **průchoda*. Mezi těmito dvěma póly leželo dalších pět slov: *sýr, rybník, týl, dvorek, pokojík*. Tvary Gsg těchto slov jsou zastoupeny v korpusu SYN2005 v různých proporcích (viz tabulku 2).

Tabulka 2. Slova v dotazníku

| tvar na -u | dokladů | proporce | tvar na -a | dokladů | proporce |
|-------------------|----------------|-----------------|-------------------|----------------|-----------------|
| 1 lesu | 0 | 0,00 % | lesa | 4316 | 100,00 % |
| 2 sýru | 118 | 9,10 % | sýra | 1185 | 90,90 % |
| 3 rybníku | 125 | 11,90 % | rybníka | 929 | 88,10 % |
| 4 týlu | 115 | 34,60 % | týla | 217 | 65,40 % |
| 5 dvorku | 150 | 74,60 % | dvorka | 51 | 25,40 % |
| 6 pokojíku | 199 | 93,00 % | pokojíka | 15 | 7,00 % |
| 7 průchodu | 194 | 100,00 % | průchoda | 0 | 0,00 % |

Při výběru slov jsem se řídil několika zásadami. Vybrané lexémy musely mít vyšší frekvenci (>150 textových slov v SYN2005), aby poměr formantů byl dostatečně spolehlivý. Musela se vyskytnout možnost lokálního pojmání, abych otestoval slova ve stejných typech kontextů a aby kontexty zněly přirozeně, a proto jsem se vyhýbal abstraktním lexémům. Vyhledal jsem slova, u kterých je málo příležitostí pro „lexikální účinky“, tj. rozdíly způsobené odlišnými významy (s výjimkou slova *týl*, kde se žádný účinek nenaskytl). Navíc jsem musel vyřadit slova, u kterých byla data zkreslena „korpusovými účinky“, např. opakovaný výskyt jednoho výrazu z novinových reklam.

Vzhledem k tvrzení v PMČ o možnosti kontextuálních rozdílů ve volbě formantů jsem vyčlenil pět syntaktických kontextů charakteristických pro Gsg:

⁴ Tentokrát jsou propria vyřazena z počtu. Tagger nerozpoznával některá řídká jména živ. rodu, vyřazení těchto jmen ale jenom zesílí zmiňovanou tendenci (zvýrazní se rozdíl mezi počtem lexémů a počtem textových slov).

Inherence (vlastnictví, charakteristika): Bydlel v novém domě na kraji <lesa>.

Adverbiální fráze pohybu (bez přídavného jména): Pustila se do housky a <sýra>.

Adverbiální fráze pohybu (+ 1 přídavné jméno): Z oploceného <dvorku> záhadně zmizela tři nákladní auta.

Jiné „adverbiální“ konstrukce s předložkou: Auto skončilo uprostřed <rybníka>.

Objektový genitiv: Nevšimla jsem si <průchodu> v hradbách a tak jsem musela obejít celý hrad.

Dotazník měl následně sedm lexémů s dvěma formanty v pěti kontextech, tj. celkem 70 vět.⁵ Až na několik výjimek pocházely všechny příklady z korpusu a byly představeny buď v úplném znění, nebo zjednodušeně (občas bylo žádoucí vypustit zbytečné detaily a zredukovat věty do kratší podoby). Respondenti měli hodnotit každou větu od 1-7 (1 = nejlepší, 7 = nejhorší). Přesný deskriptor měly jenom krajní body.

Dotazníky tohoto typu nejčastěji měří **aktivní produkci** tvarů, tj. nabízí se mezera, do které má respondent doplnit vhodný tvar (typický příklad je uveden v Kasalovi (1992)). Můj dotazník se však opíral na hodnocení přijatelnosti variant, tj. na **pasivní recepci** tvarů, a to z několika důvodů. Reprezentativnost korpusu SYN2005 znamená, že korpus představuje průřez psaného světa, který obklopuje průměrného českého čtenáře, ale tento fakt nás nepřivede k závěrům o produkčních schopnostech jednotlivců. Jinak řečeno, skutečnost, že čtenář vidí častěji nebo řidčeji v psaných textech ten či onen tvar neimplikuje nutně, že aktivně ovládá a užívá daný tvar ve stejné míře, nýbrž že by ho měl do větší nebo menší míry akceptovat (viz hypotézy výše). Dalším důvodem bylo, že vyplnění mezer předpokládá odpověď „buď/nebo“, zvýrazňuje jenom nejběžnější variantu a neposkytuje možnost rozlišení míry přijatelnosti. Proto byl pro mě tento přístup málo vhodný.

Vzhledem k délce dotazníku jsem testoval každý formant na jiné větě, aby respondenti nepodlehli nudě kvůli opakovaným replikám. Proto jsem vytvořil dvě základní mutace dotazníku: v první verzi se formant *-a* vyskytuje v jedné skupině vět, v druhé verzi se vyskytuje v jiné skupině. Tím se omezilo riziko vedlejších účinků vyplývajících z jednotlivých vět. Kromě toho existovaly i dvě mutace pořadí: v druhé jsem změnil pořadí vět, abych se vyhnul riziku, že bude volba tvarů pořadím ovlivněna. První čtyři věty každé mutace však vždy představovaly dvě výpovědi hodnocené jako velmi normální a přijatelné a dvě jako naprosto nepřijatelné. Tím jsem chtěl nabádat respondenty, aby používali celou škálu možných odpovědí.

Celkem byly tedy vytvořeny čtyři mutace dotazníku nahodile distribuované mezi respondenty, kteří dotazník vyplnili v mé přítomnosti, a to buď jednotlivě anebo v malých skupinkách.

Pro lepší statistické zpracování jsem zvolil škálu s přesnými body. Zvážil jsem nejdříve přístup tzv. odhadnutí rozsahu – *magnitude estimation* – ale nakonec jsem ho nepoužil, protože tento přístup je pro respondenty mnohem časově náročnější a únavné. Dotazník byl i tak poměrně dlouhý, a proto jsem nepoužil tzv. rozptylovací věty (*distracter sentences*). Slovo, o které se jednalo, bylo označeno tučným písmem, abych se vyhnul zavádějícím odpovědím (např. upoutá pozornost respondenta další jev ve větě a hodnotí ji podle toho).⁶

⁵ Na seminářích o výsledcích dotazníku se komentář k volbě slov nejčastěji zaměřoval buď k faktu, že je jich málo, nebo k jednotlivým charakteristikám vybraných lexémů. Není pochyb, že by bylo žádoucí zahrnout do dotazníkové akce více lexémů, jenže s narůstajícím počtem slov se musí pak omezit data o každém slově, což ubírá na úplnosti přehledu. Výběr slov bude o něco rozšířen v dalších fázích projektu.

⁶ Děkuji dr. Evě Dąbrowské a dr. Dagmar Divjak za odborné rady ohledně sestavení dotazníku a mgr. Luďku Knittlovi za jazykovou redakci dotazníku. Řídil jsem se pokud možno zásadami v příručkách o dotazníkových akcích, např. Schütze (1996) a Cowart (1997). Jsem si vědom, že se v syntaktických dotaznících doporučuje neobracet pozornost respondentů na jevy a tuto zásadu jsem tedy porušil v absenci rozptylovacích vět a v označení studovaného tvaru. Domnívám se ale, že si v morfologické studii můžeme dovolit větší volnost, než ve studii syntaktické: morfologický tvar je jasně vyhraněná věc a jeho označení nenutí respondenta do

4. Charakteristika odpovědí

Pilotní studie, která se uskutečnila v Sheffieldu během ledna a února r. 2008, měla ovšem některé rysy, na které je žádoucí upozornit.

Počet respondentů (N=20) byl poměrně nízký. Mezi nimi byli i někteří studenti českého jazyka a literatury, kteří mají vysokou úroveň sebereflexe ohledně jazyka. Čtyři respondenti emigrovali před více než 10 lety, i když v daném případě možnost vlivu angličtiny jako druhého jazyka je díky absenci podobných gramatických jevů minimální.

Všichni respondenti mě znali osobně, a proto se i přes délku dotazníku zjevně snažili: všechny vyplněné dotazníky byly použitelné. Nicméně někteří respondenti nepoužili celou škálu: vyhnuli se sedmému bodu, či použili jenom body 1 a 7.

Průměrný věk mých respondentů byl 29, středový věk 25. Nejmladšímu bylo 21, nejstaršímu 67. Z nich bylo 14 z Čech, 4 z Moravy a 2 ze Slezska, což je blízko proporcí v českém obyvatelstvu. Nebyly nalezeny žádné významné rozdíly mezi mediány či průměry skupin, které odpovídaly na různé mutace dotazníku; opatření proti nechtěným vedlejším účinkům byla tedy buď úspěšná či zbytečná.

5. Statistická analýza dotazníku⁷

Provedli jsme test ANOVA („analýza proměnných“) na kombinovaná skóre pro každé slovo (hodnocení přijatelnosti *-a* minus hodnocení přijatelnosti *-u*). Model měl dva faktory: slovo (7 úrovní) a syntaktická konstrukce (5 úrovní). Hlavní účinek má slovo:

$$F(6, 114) = 36,27, p < 0,001, \text{ částečná } \eta^2 = 0,66$$

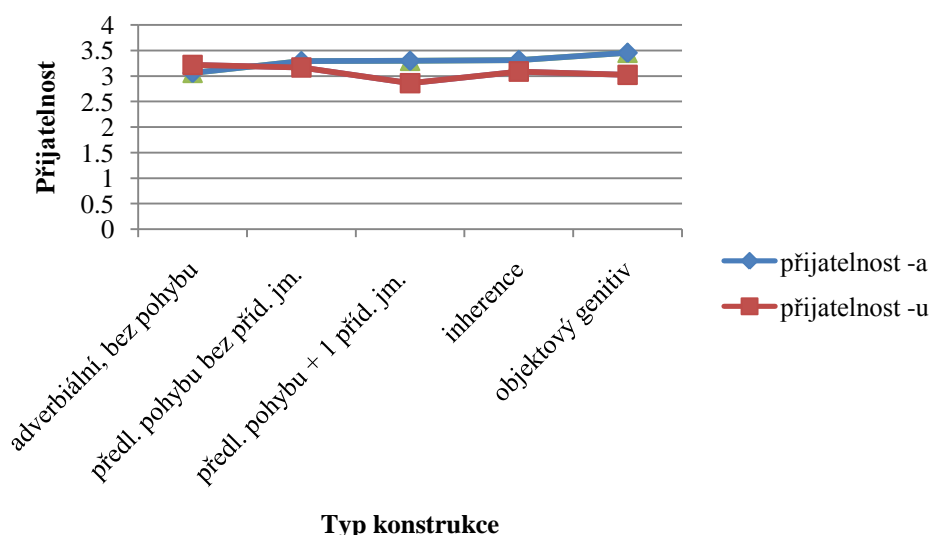
Potvrdila se existence interakce mezi syntaktickou konstrukcí a slovem:

$$F(24, 456) = 2,45, p < 0,001, \text{ částečná } \eta^2 = 0,15$$

Z toho vyplývá, že volba slov je velmi významný faktor v hodnocení přijatelnosti toho či onoho formantu. Za povšimnutí stojí i účinek jednotlivých konstrukcí u některých slov, výsledek však není tak významný.

Testy korelace ukázaly, že mezi syntaktickým kontextem a přijatelností je korelace s nízkou významností (-0,53). Výsledky jsou v tabulce 3.

Tabulka 3. Syntaktické konstrukce



přemýšlení o gramatických strukturách, o kterých nic neví. To, co ztratíme na bezprostřednosti odpovědi, získáme zpátky na přesnosti jejího zaměření.

⁷ Za statistické zpracování dat a pomoc s jejich interpretací děkuji dr. Ewě Dąbrowské.

Nakonec jsme provedli t-testy (testy významnosti), které měří rozdíly v přijatelnosti mezi jednotlivými páry slov. Použili jsme kombinovaná skóre jako u testu ANOVA a všimli jsme si významných rozdílů mezi skóre pro některá sousední slova:

les a *sýr*, $t(19) = 4.35$, $p < 0,001$

pokojík a *průchod*, $t(19) = 7.95$, $p < 0,01$

Mezi ostatními sousedními slovy jsme žádné významné rozdíly nenašli. Naopak u některých nesousedních párů byly rozdíly statisticky významné:

les a *rybník*, $t(19) = 6,23$, $p < 0,001$

rybník a *dvorek*, $t(19) = 5,58$, $p < 0,001$

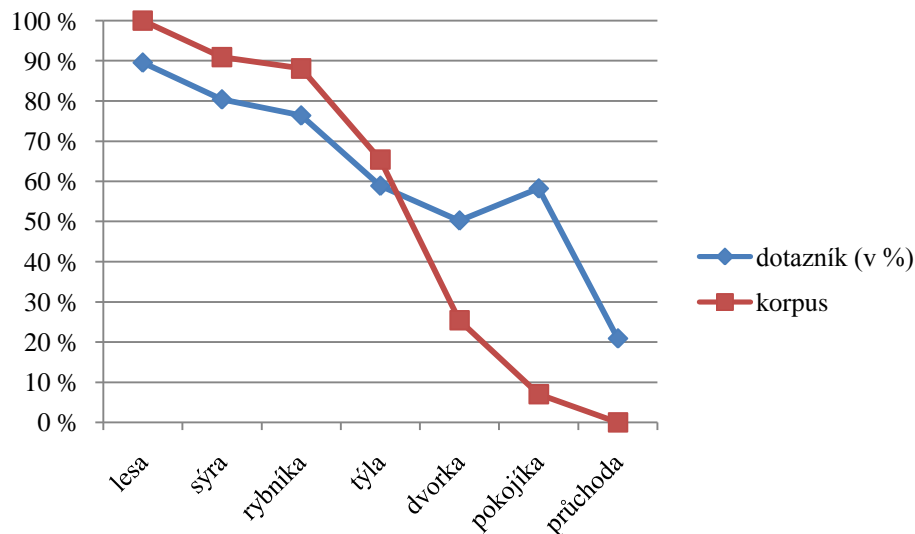
dvorek a *průchod*, $t(19) = 7,01$, $p < 0,001$

Je důležité poznamenat, že ve čtyřech z pěti případů jde o srovnání krajních slov (*les*, *průchod*) s prostředními, tj. nejde ve skutečnosti o korelace mezi slovy s evidovanou variací ve tvarech.

6. Statistické srovnání dat z dotazníku s daty z korpusu

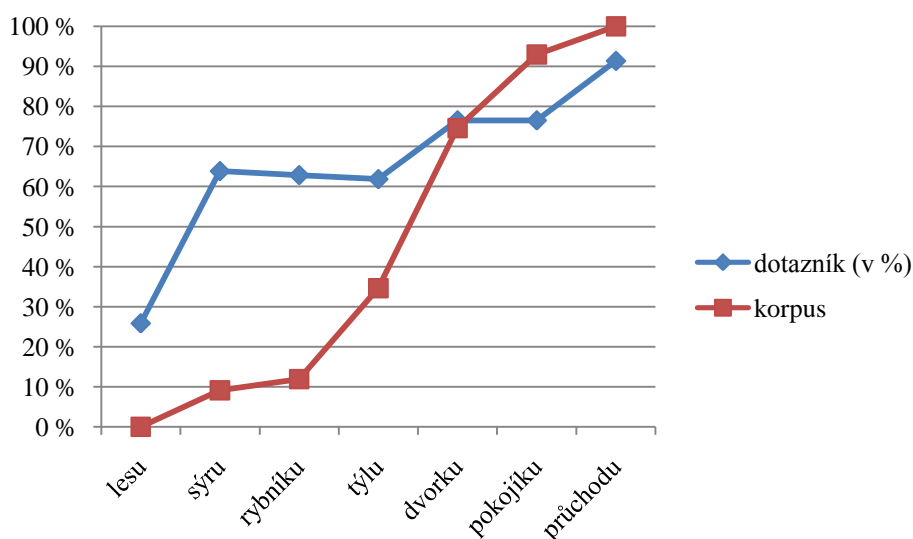
V další fázi analýzy jsme srovnali korpusovou frekvenci s hodnocením přijatelnosti z dotazníku. Výhradně pro tento účel jsme konvertovali hodnocení přijatelnosti ze škály 1–7 na procenta 0 % – 100 %. S formantem *-a* byla korelace 0,88 s krajními slovy (*les*, *průchod*), 0,83 bez krajních slov, tj. velmi významná – viz tabulku 4:

Tabulka 4. Formant *-a*: korelace mezi korpusovými daty a údaji z korpusu



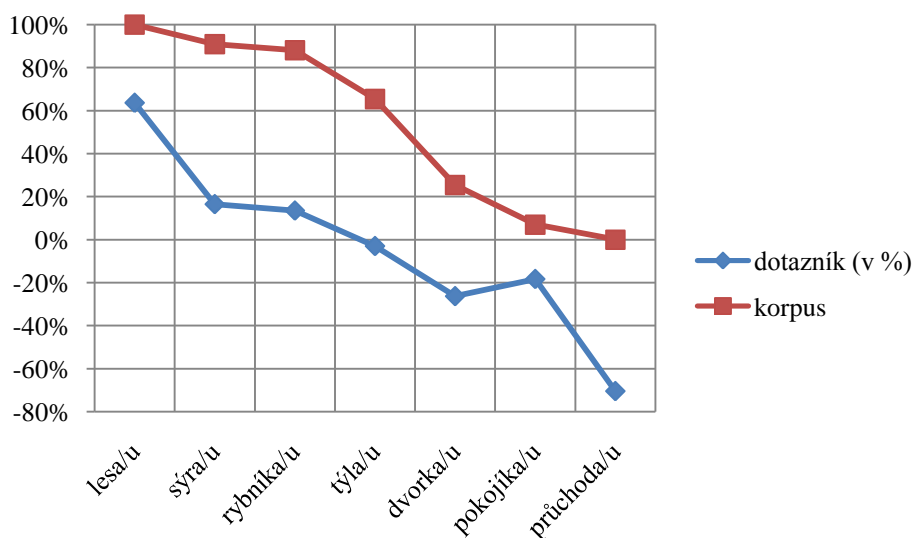
S formantem *-u* jsme našli korelaci 0,81 s krajními slovy, 0,92 bez krajních slov, opět byla korelace velmi významná (viz tabulku 5):

Tabulka 5. Formant *-u*: korelace mezi korpusovými daty a údaji z korpusu



S kombinovanými skóre se vyskytla korelace 0,89 s krajními slovy, 0,94 bez krajních slov (velmi významná), viz tabulku 6:

Tabulka 6. Kombinovaná skóre: korelace mezi korpusovými daty a údaji z korpusu



7. Závěry: korpusová data a přijatelnost

Výsledky pilotní studie potvrzují výzkum o syntaxi, viz např. Divjak (2008): Korpusová data poukazují na významnou korelaci s daty o přijatelnosti tvarů, vztah mezi korpusovou frekvencí a přijatelností ale není jednoduchý. Z korpusových dat nelze určit úroveň přijatelnosti nějakého tvaru, obzvlášť pro tvary s nízkou frekvencí.

Divjak poukazuje na to, že nízká frekvence syntaktické konstrukce v korpusu neodpovídá nepřijatelnosti pro rodilé mluvčí, a data z této pilotní studie potvrzují její závěry i pro morfologii. Naopak, opět podle Divjak, se ale zdá, že nízká přijatelnost (**průchoda*, **lesu*) souvisí s nízkou frekvencí. Vysoká frekvence (>50 %) svědčí o vysoké přijatelnosti, avšak vyšší přijatelnost (<3,5) nesevědí o vysoké frekvenci. Podle našich dat tvary s nízkou

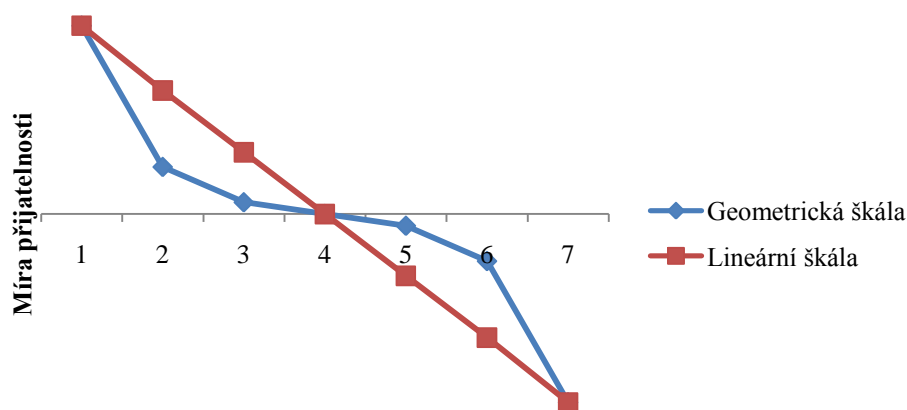
frekvencí (*sýru, rybníku, pokojíka* 10 %) mohou být pro rodilé mluvčí stejně přijatelné jako tvary s vyšší frekvencí (*týla* 65 %, *týlu* 35 %, *dvorka* 25 %).

Pro odhad přijatelnosti korpusová frekvence tedy sama nestačí. Musíme ji zkombinovat s pečlivou analýzou empirických dat. Hypotéza 1 proto není potvrzena; úroveň přijatelnosti není pevně vázána na korpusovou frekvenci. Hypotéza 3 ale také není pravdivá, protože se potvrdila určitá souvislost (korelace) mezi korpusovými daty a posouzením přijatelnosti. Na základě těchto dat se jeví jako nejpravděpodobnější předpoklad Hypotézy 2: obecná korelace existuje, nesmí se ale přeceňovat významnost procentuálního zastoupení v korpusu vzhledem k určení přijatelnosti. Významné pro určení přijatelnosti se zdají být jenom frekvence nad 60 % a možná i totální absence nějakého tvaru.

8. První dovětek: Jsou škály přijatelnosti lineární či geometrické?

Je na místě se zeptat, jestli některé efekty zmíněné výše mohou vzniknout jako vedlejší účinek výzkumných metod. Sedmibodová škála má být lineární, tj. v zásadě rozdíl v přijatelnosti mezi 1 a 2 je stejně významný jako rozdíl mezi 2 a 3, 3 a 4 atd. Výrazné definice krajního bodu („jediné možné“, „nepřijatelné“) a neurčitost definicí bodů mezi dvěma extrémy („více“ či „méně“ (ne)přijatelné) může spíše svědčit o existenci geometrické škály, při které respondenti vnímají významnější rozdíl mezi krajním bodem a sousedním oproti méně patrnému rozdílu mezi prostředními sousedními body (viz tabulku 7).

Tabulka 7. Lineární a geometrické škály



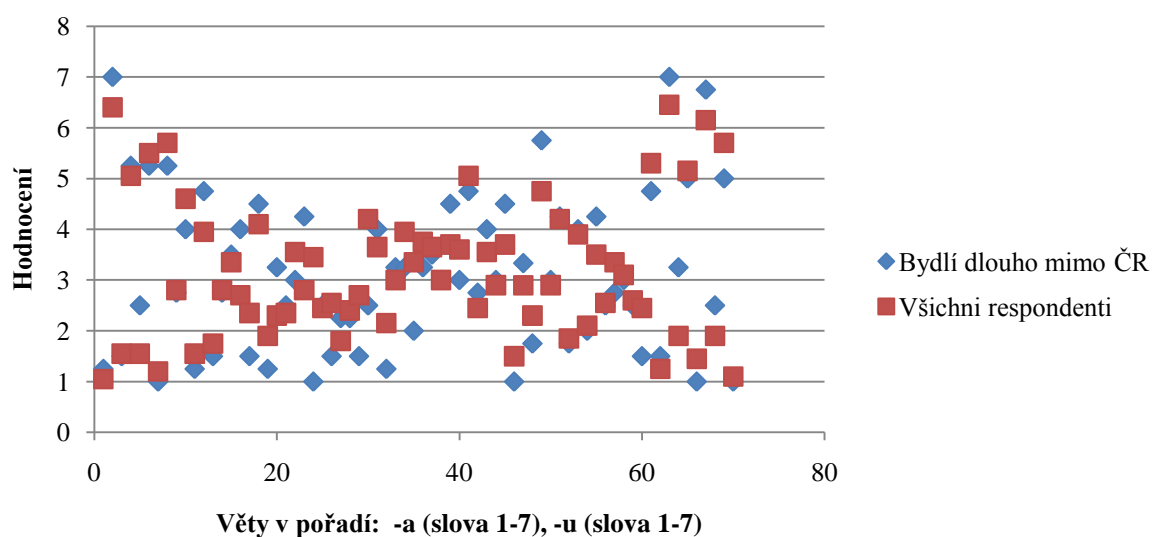
1 = jediný možný tvar, 7 = nepřijatelný tvar

Domnívám se, že takovou možnost musíme akceptovat jako potenciální nedostatek metody hodnocení přijatelnosti: potvrzení významnosti výsledků nám dovoluje, abychom se spoléhali na rozlišovací schopnosti respondentů.

9. Druhý dovětek: Dlouhodobý pobyt mimo ČR

Mezi mými respondenty byli čtyři Češi dlouhodobě žijící mimo Českou republiku. Z dat se zdá, že dlouhodobí emigranti odmítají některé tvary kategoričtěji než mluvčí s trvalým bydlištěm v ČR. V odpovědích v tabulce 8 vidíme více „krajních“ hodnot (blíže k 1 a k 7), než tomu tak je pro všechny respondenty. Je tedy možné, že emigranti jsou více „sami“ se svou vlastní češtinou, nejsou natolik vystaveni rozmanitosti současného jazyka, a proto se přiklánějí k razantnějším rozhodnutím podle toho, mají-li daný tvar ve vlastním repertoáru. Tento efekt ale není pravidelný a nezdá se mít žádný obecný vliv na jejich hodnocení tvarů.

Tabulka 8. Dlouhodobý pobyt mimo ČR



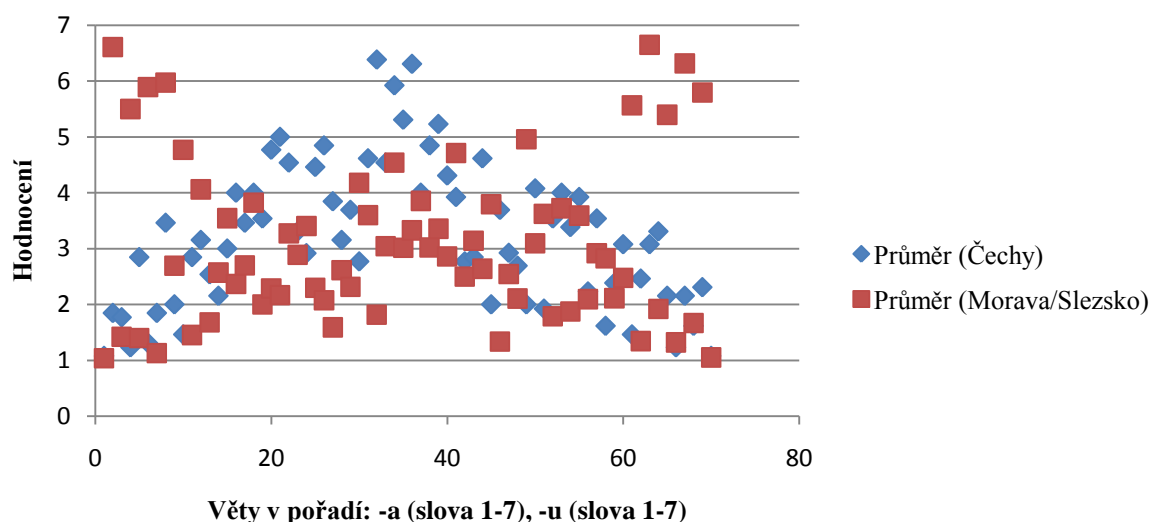
Dlouhodobý pobyt v zahraničí se nezdá mít žádný konzistentní vliv na preference o užití pádových tvarů (i když mé výsledky i existence literatury o změnách v syntaxi a lexikonu emigrantů naznačují potřebu detailnějšího výzkumu tohoto jevu).

10. Třetí dovětek: Regionální rozdíly uvnitř ČR

Pádové tvary v češtině mohou mít značnou variabilitu podle regionu, kde mluvčí bydlí a odkud pochází. Respondenti pilotní studie tvořili v hrubých rysech (Čechy – Morava – Slezsko) reprezentativní zastoupení českého obyvatelstva, a proto bylo možné aspoň položit otázku, zda existují nějaké patrné rozdíly mezi preferencemi Čechů a Moravanů/Slezanů. U jiných jevů se často tvrdí, že moravské tvary jsou konzervativnější, a podle toho bychom čekali, že Moravané budou upřednostňovat tvary na *-a*, které jsou pro některá slova starší a pro jiná představují extensi v češtině méně produktivního formantu.

Rozdíly se však ukázaly být nepatrné. Průměrné moravské a slezské hodnocení tvary na *-a* je 3,02 versus 3,16 pro Čechy. Co se týče hodnocení tvarů na *-u*, průměrné moravské a slezské hodnocení je 3,18 versus 3,04 pro Čechy (viz tabulku 9).

Tabulka 9. Rozdíly v jednotlivých odpovědích mezi Čechy a Moravany/Slezany



Regionální rozdíly nebyly statisticky významné. Jak je vidět z tabulky 9, rozložení odpovědí je však zajímavé a zasloužilo by další pozornost s větším počtem respondentů a slov. Zdá se, že přes nevýznamnost dat mohou existovat regionální rozdíly v preferencích týkajících se jednotlivých slov.

11. Závěry

Data z velkého reprezentativního korpusu se zdají souviset s odhady přijatelnosti rodilých mluvčích do určité míry, ale ne zcela. Velké procentuální zastoupení tvarů v korpusu svědčí o vysoké přijatelnosti, ale zastoupení pod 50 procent se neukázalo jako spolehlivé měřítko přijatelnosti. Tato zjištění se shodují s existující literaturou o syntaktických jevech a po ověření na větších skupinách a rozsáhlejší slovní zásobě mohou být užitečná pro další analýzy korpusových dat.

Zdroje

- BERMEL, N. (1993): Sémantické rozdíly v tvarech českého lokálu. *Naše řeč* 76, s. 192-198.
- BERMEL, N. (2004): *V korpuse nebo v korpusu? Co nám řekne (a neřekne) ČNK o morfologické variaci v tvarech lokálu.* In Hladková, Z. – Karlík, P (eds.), *Čeština – univerzália a specifika 5.* Praha: Nakladatelství Lidové Noviny, s. 163-171.
- COWART, W. (1997): *Experimental Syntax: Applying Objective Methods to Sentence Judgments.* Thousand Oaks, CA: Sage Publications.
- CUMMINS, G. (1995): Locative in Czech: -u or -e: Choosing locative singular endings in Czech nouns. *Slavic and East European Journal* 39, s. 241-260.
- ČNK: *Český národní korpus – SYN2005.* Ústav českého národního korpusu FF UK. Dostupné na webu: www.korpus.cz
- DIVJAK, D. (2008): On (in)frequency and (un)acceptability. In: Barbara Lewandowska-Tomaszczyk (ed.): *Corpus Linguistics, Computer Tools and Applications - State of the Art.* Frankfurt am Main: Peter Lang, s. 213–233.
- HALLIDAY, M. A. K. (1991a): Corpus studies and probabilistic grammar. In: Aijmer, K. and Altenberg, B. (ed.): *English Corpus Linguistics.* New York a Londýn: Longman, s. 30–43.
- HALLIDAY, M. A. K. (1991b): Towards probabilistic interpretations. In: Ventola, E. (ed.): *Functional and Systemic Linguistics: Approaches and Uses.* Berlín a New York: Mouton de Gruyter, s. 39–61.
- HALLIDAY, M. A. K. (1992): Language as system and language as instance. In: Svartvig, J. (ed.): *The Corpus as a Theoretical Construct: Directions in Corpus Linguistics.* Berlin and New York: Mouton de Gruyter, s. 61–77.
- KASAL, J. (1992): Dublety a jejich užití. In: *Philologica* 65, s. 107-114. Olomouc: Univerzita Palackého.
- KEMPEN, G. – HARBUSCH, K. (2005) The relationship between grammaticality ratings and corpus frequencies: A case study into word order variability in the midfield of German clauses. In: Kepsler, S. – Reis, M. (eds.): *Linguistic Evidence: Empirical, Theoretical and Computational Perspectives.* Berlín a New York: Mouton de Gruyter, s. 329–349.
- KLIMEŠ, L. (1953): Lokál singuláru a plurálu vzoru „hrad“ a „město“. *Naše řeč* 36, s. 212–219
- KRÁLÍK, J. (2001): Vyvážení zdrojů Synchronního korpusu češtiny SYN2000. *Slovo a slovesnost* 62, s. 38-53.
- MČ: *Mluvnice češtiny.* Petr, J. (ed.), Praha: Academia, 1986.
- OLIVA, K. – DOLEŽALOVÁ, D. (2004): O korpusu jako o zdroji jazykových dat. In: Karlík, P. (ed.): *Korpus jako zdroj dat o češtině.* Brno: Masarykova univerzita, s. 7–10.

- PMČ: *Příruční mluvnice češtiny*. Karlík, P. – Nekula, M. – Rusínová, Z. (eds.), Praha: Nakladatelství Lidové Noviny, 1995.
- RUSÍNOVÁ, Z. (1992): Některé aspekty distribuce alomorfů (genitiv a lokál sg. maskulin). *Sborník prací filozofické fakulty Brněnské univerzity, Řada A (Jazykovědná)* 40, s. 23-31.
- SEDLÁČEK, M. (1982): V Záhřebě i v Záhřebu. *Naše řeč* 65, s. 11-15.
- SCHÜTZE, C. (1996): *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Chicago: University of Chicago Press.
- ŠTÍCHA, F. (2008??).
- ŠULC, M. (2001): Tematická reprezentativnost korpusů. *Slovo a slovesnost* 62, s. 53-61.