
Article:
Bermel, N.H., Knittl, L. and Russell, J. (2014) Absolutní a proporcionální frekvence v ČNK ve světle výzkumu morfosyntaktické variace v češtině. Naše řeč, 97 (4-5). 216 - 227 .

Absolutní a proporcionální frekvence v ČNK ve světle výzkumu morfosyntaktické variace v češtině[1]

Neil Bermel, Luděk Knittl, Jean Russell

Abstract:

This contribution discusses three ways of operationalising the notion of frequency as it relates to how often an item occurs in a corpus: the proportional frequency of forms (i.e. percentage of time that one or another variant is found) and two ways of looking at absolute frequency. Working with data from unmotivated morphological variation in Czech case forms, we show that different types of data contribute to some extent to the way variation is perceived and implemented by native speakers, but suggest that proportional frequency seems most salient for speakers in forming their impressions and shaping their behaviour.

Klíčová slova: korpusová lingvistika, frekvence, tvarosloví, empirický výzkum, dotazníky, čeština / corpus linguistics, frequency, morphology, empirical research, questionnaires, Czech

1. Úvod

Lingvisté, kteří pracují s korpusem, už jsou dávno zvyklí uvádět frekvenční údaje: kolikrát se vyskytl ten nebo onen jev v daném korpusu. Četnost v korpusu citujeme, protože předpokládáme, že korpus něco zastupuje – např. svět textů – a následkem toho jsme z frekvencí v korpusu schopni o tomto světě textů něco vyčíst. Jinak řečeno, pomocí korpusu probíhá jistá operacionalizace našich otázek o jazyce.

V empirickém výzkumu je způsob operacionalizace otázka prvořadého zájmu: pomocí operacionalizace manipulujeme výzkumnou otázku, vytváříme z ní měřitelnou verzi naší hypotézy, kterou lze pomocí dat potvrdit, nebo vyvrátit. My, lingvisté, ale často považujeme tento krok za implicitní a neupřesňujeme ho, ani pro čtenáře, ani pro sebe. Cílem tohoto příspěvku bude tedy vyjasnit způsoby operacionalizace otázek a jejich vliv na naše výsledky a

[1] Tento článek vznikl v rámci projektu „Acceptability and forced-choice judgements in the study of linguistic variation“ s podporou nadace Leverhulme Trust (č. RPG-407).

závěry. Zaměříme se na různé interpretace termínu frekvence ve velkých korpusech, a to na základě výzkumu „konkurujících si“ morfologických variant. Vyjasníme způsoby, jak můžeme v korpusu počítat frekvenci těchto tvarů a jak tyto počty souvisí s dalšími empirickými daty o hodnocení a užití jazyka. Na základě těchto sond navrhneme, který typ frekvence se nám zdá být ve výzkumu morfologické variace nejužitečnější.

2. Výzkumný problém

Začněme jednoduchým příkladem morfologické variace analogickým tomu, který je zmíněn mj. v pracích Čech (2012), Cvrček – Kodýtek (2013). Data z velkého reprezentativního korpusu SYN2010[2] o výskytu variantních tvarů v Lpl ž. rodu (vzory *píseň, kost*) můžeme prezentovat třemi způsoby. První způsob je podle počtů dokladů jednotlivých tvarů:

(1) Doklady tvarů v Lpl v korpusu SYN2010

nocích (427), *pamětech* (268), nemocech (108), *nemocích* (24), *pěstích* (9), nocech (5), *pěstech* (4), *pamětích* (3).[3]

Tato data můžeme prezentovat také v morfologických opozicích, jelikož morfologická variace předpokládá, že se v určité situaci objeví jenom jedna z velmi omezeného počtu variant (zprav. ne více než tři).:

(2) *Doklady tvarů v* Lpl v *morfologických opozicích*

pamětech (268)	<>	*pamětích* (3)
nemocech (108)	<>	*nemocích* (24)
pěstech (4)	<>	*pěstích* (9)
nocech (5)	<>	*nocích* (427)

Přepracováním tabulky (2) můžeme přiřadit výsledky podle procent zvolených variant:

(3) *Doklady tvarů v* Lpl v *morfologických opozicích podle procent*

pamětech (98,9 %)	<>	*pamětích* (1,1 %)
nemocech (81,8 %)	<>	*nemocích* (18,2 %)
pěstech (30,8 %)	<>	*pěstích* (69,2 %)
nocech (1,2 %)	<>	*nocích* (98,8 %)

[2] O reprezentativnost korpusů ČNK viz např. Čermák – Králík – Kučera (1997), Králík – Šulc (2005).

[3] Korpusová lingvistika často operuje s pojmem *relativní frekvence,* tj. výskyt v „standardním“ korpusu s milionem slov. V případě reprezentativních složek ČNK bychom museli dělit tyto frekvence stem: RF *nocích* je tedy 4,27.

Tabulka (3) vypadá přehledněji: podle ní korpus jasně dává přednost tvarům *pamětech, nemocech, pěstích, nocích* oproti tvarům *pamětích, nemocích, pěstech, nocech.* Nahrazením počtů pouhými procenty jsme ale ztratili určité informace. V tabulce (1) si všimněme, že dokladů většinového tvaru *pěstích* je mnohem méně než dokladů menšinového tvaru *nemocích*: bylo by podložené tedy doopravdy předpokládat, že *nemocích* je dispreferovaný tvar, když je četnější než údajně preferovaný *pěstích*?[4]

Každý z těchto přístupů zachycuje alespoň jednu důležitou skutečnost o datech a nebere v úvahu další. Zařazení v tabulce (1) nabízí data generovaná korpusem: podle něj měříme všechny tvary stejným měřítkem a nezasahujeme do gramatiky, tj. nepokládáme za relevantní, které tvary „si konkurují" a do kterého vzoru (*píseň* či kost) daný tvar patří. Zařazení v tabulce (3) zasazuje korpusové údaje hned do předem předpokládaných jazykových struktur, které slouží jako základ analýzy. Tímto přístupem upřednostňujeme vztahy mezi tvary v jedné morfologické „buňce" a umožňujeme i srovnání mezi opozicemi jako celkem. Mezní pozici má tabulka (2), která nabízí srovnání těchto tvarů podle morfologické „konkurence", ale podává korpusová data v jejich původní formě.

V dalších částech se zaměříme na otázku: jestliže můžeme operacionalizovat korpusovou frekvenci dvěma způsoby (tj. absolutními hodnotami a proporcionálními hodnotami), který z nich úžeji souvisí s chováním uživatelů jazyka (měřeným hodnoceními přijatelnosti a doplňováním tvarů)?

3. Širší kontext zkoumané problematiky: absolutní a proporcionální frekvence

Jak v analýze dat, tak v popisu jejích výsledků se setkáváme s potřebou členit data do pásem či kategorií. Některé typy statistické analýzy (např. analýza rozptylu nebo chí-kvadrát) vyžadují místo seřazení dat na škále jejich členění do jasných frekvenčních kategorií. Toto seskupování výsledků zároveň může pomoci čtenářům v rozluštění relevance či významu frekvence jako obecného jevu.

V našem případě jde o seřazení dat do různých „pásem" jak pro absolutní, tak pro proporcionální frekvenci. Začínáme operacionalizací pojmu *absolutní frekvence*.

V korpusové lingvistice neexistují pevně stanovené hranice pro vysokou a nízkou absolutní frekvenci, a to proto, že obecná četnost jevů bývá různá.

Bybee (2007:16) např. pracuje hlavně s absolutními frekvencemi a doporučuje určit pro každý jev jinou hranici mezi pásmy. Její kritéria jsou: 1/ existence „frekvenční mezery", která

[4] Např. Čech (2012: 210–211) poukazuje na to, že používání procentuálního zastoupení tvarů bez ohledu na jejich skutečné frekvence může být zavádějící – významnost je podle něj úzce spjata s velikostí vzorku.

rozdělí škálu do dvou části, přičemž 2/ každá část obsahuje 30 až 70 procent lemmat. V jedné studii elize /t/ a /d/ v angličtině (2002: 264) přijala jako hranici frekvenci 35 tvarů v milionovém korpusu, v další studii elize /d/ a /ð/ ve španělštině (2002: 265–266) stanovila hranici 100, přičemž použila korpus o velikosti 1,1m slov. Ve třetí studii (Bybee – Eddington 2006: 329) s daty ze dvou korpusů s celkovým rozsahem 2 mil. slov měl jeden typ spojení vysokou frekvenci se 17 výskyty a ostatní spojení měla nízkou frekvencí s 9 výskyty nebo méně.

Data, s kterými operujeme dále, pocházejí ze vzoru hrad a týkají se variantnosti v Gsg a Lsg (typu jazyka/jazyku, *hradě/hradu*). Naše korpusová data čerpají ze SYN2010, který má něco přes 100m textových slov a je 50krát až stokrát větší než korpusy použité Bybeeovou.[5] Abychom navázali na proporce využité v jejích studiích, hranice mezi „vysokou" a „nízkou" frekvencí mohou být v našem stomilionovém korpusu kdekoliv mezi 850 a 9 090 tokeny.

Jevy, které jsme testovali, jsou méně frekventované, zvlášť bereme-li v potaz, že každé české slovo má sedm pádů s formálním odrazem v morfologické rovině. Jejich rozložení v korpusu odpovídá Zipfovu zákonu, tj. je malé množství lexémů s vysokou frekvencí a valná většina slov s kolísáním má frekvenci minimální (v případě Gsg muž. rodu typu jazyka/jazyku je ze 112 lexémů až 52 s méně než 100 dokladů a v případě Lsg muž. rodu typu *hradě/hradu* je z 391 lexémů až 186 s méně než 100 dokladů). Kvůli nespolehlivosti dat s nižšími frekvencemi jsme se rozhodli testovat pouze lexémy s četností nad 100 dokladů v korpusu.[6]

V obou pádech existuje „frekvenční mezera" (viz Bybee 2007 výše) kolem hranice 1000 dokladů. Tato mezera rozděluje lexémy do dvou nerovnoměrných skupin. Ve skupině s vysokou frekvencí zůstalo jenom 52 z 391 Lsg tvarů (13,3 %) a 23 ze 112 Gsg tvarů (20,5 %). Mohli jsme samozřejmě hledat nižší hranici, aby skupiny byly vyvážené, ale rozhodli jsme se zůstat u hranice 1000 výskytů v stomilionovém korpusu. Hlavní důvod byl, že termín „vysoká frekvence" skoro nikde neznamená méně než 7–8 v milionovém korpusu (700–800 v stomilionovém) a chtěli jsme zachovat možnost širší relevance našich výsledků.

Druhá část této operacionalizace se týká frekvence *proporcionální*. V korpusově založených studiích je využívána často. Stejně jako u absolutní frekvence se i tady setkáváme s různými přístupy. Halliday (1992: 65–66) navrhuje hranice 9:1 a 1:9, mezi kterými vnímáme jednu variantu jako „běžnou" a druhou jako „výjimečnou"; v případech s méně odlišnými proporcemi (např. 4:1, 2:1, 3:2) jde podle něj o varianty se stylovým, významovým, či jiným

[5] O celkové frekvenci těchto konkurující si koncovek viz Bermel – Knittl 2012a: 249

[6] Tím jsme se zároveň vyhnuli dalšímu problému diskutovanému u Čecha (2012: 211) a Cvrčka – Kodýtka (2013: 141) o významnosti malých výčtů.

funkčním rozdílem. Hare a kol. (2001) navrhuje jiné členění, a to do tří pásem s hranicemi 1:2 a 2:1. Jiné systémy předložené pro češtinu uvádíme v tabulce (4):

(4) *Členění dokladů morfologické variace do pásma*

Zdroj: Bermel − Knittl, 2012

Cíl: Frekvenční pásma pro morfologické opozice (podle „ustupující“ koncovky {a}, {ě})

Záměr: Popis empirických výsledků přijatelnosti ve vztahu k původním korpusovým datům

izolované	*příznakové*	*menšinové*	*rovnocenné*	*většinové*	*bezpříznakové*	*dominantní*
pod 1 %	1–9 %	10–29 %	30–69 %	70–89 %	90–99%	nad 99 %

Zdroj: Hebal-Jezierska, 2007

Cíl: Frekvenční pásma pro každou variantu (Npl. {i}, {ové}, {é})

Záměr: Seskupení variant podle užití v kontextu

sporadické	*variantní*	*dominantní*
0–1 %	1–14 %	15–100 %

Zdroj: Cvrček a kol., 2010, dále vysvětleno ve Cvrček – Kodýtek, 2013

Cíl: Frekvenční pásma pro každou variantu (morfologie)

Záměr: Popis korpusové frekvence v gramatice

nikdy/ skoro nikdy	*zřídka*	*někdy*	*stejně*	*často*	zpravidla	*vždycky/skoro vždycky*
0 –1%	1–10%	10 –35%	35– 65%	65–90%	90 –99%	nad 99%

Zdroj: Šimandl 2010[7]

Cíl: Frekvenční pásma pro morfologické opozice

Záměr: Popis variantnosti podle výsledků z korpusu

marginální – monopolní	*minoritní – majoritní*	*ekvipolentní*	*majoritní – minoritní*	*monopolní – marginální*
≤ 5 % > 95,0 %	5,1–39,9 % 94,5–60,1 %	40–60 % 40–60 %	60,1–94,5 % 39,9–5,1 %	> 95,0 % ≤ 5 %

Zdroj: Hebal-Jezierska − Bermel, 2011

Cíl: Frekvenční pásma pro morfologické varianty

Záměr: Vymezení pásem vzhledem k užití variant v kontextu a k výzkumu přijatelnosti variant

sporadické	*minoritní*	*majoritní*
0-2 %	2–49%	50–100%

[7] Existuje nepatrná mezera mezi *majoritní* a *monopolní* skupinou (94,5 % – 95,0 %), která není odůvodněna.

Společné mají všechny škály jenom jedno: čím rovnocennější je zastoupení obou variant v korpusu, tím širší je pásmo, které charakterizuje opozici. Jsou k tomu dva důvody, které vyplývají ze studií Hebal-Jezierské a z našich studií: první se vztahuje k výsledkům analýzy a druhý k operacionalizaci výzkumné otázky.

Co se týče výsledků, dostupná literatura nabízí dva způsoby hodnocení výsledků. Buď přistupujeme čistě na základě korpusových dat, anebo k nim přidáme i data z nekorpusových zdrojů (dotazníky atd.). Korpusová data zkoumala např. Hebal-Jezierska a nevnímala podstatné rozdíly ve fungování tvarů v tomto širokém prostředním frekvenčním pásmu: tím měla na mysli, že nedošlo k stylistickým či jiným omezením tvarů, které byly zastoupeny v korpusu s frekvencí nad 15 procent, kdežto pod touto hranicí našla řadu omezení v užívání tvarů. Větší rozkolísanost opozic je tedy vnímána tam, kde dojde k větším rozdílům v korpusové frekvenci, tedy u málo frekventovaných koncovek a u jejich protějšků, tj. tvarů, které se užívají téměř pravidelně. Výsledky hodnotících dotazníků (viz např. Bermel − Knittl 2012a, 2012b) potvrdily, že rodilí mluvčí pociťují rozdíl mezi tvary, které v životě uvidí jen málokdy a těmi, které potkávají v psaných textech pravidelněji.

Co se týče operacionalizace korpusových dat ve výzkumu: abychom měli dostatek materiálu, musíme mít v každém pásmu dost příkladových slov. Lexémy jsou však rozloženy bimodálně, tj. je jich víc v okrajových frekvenčních pásmech (proporcionální frekvence 0–10 % a 90–100 %). V prostředních pásmech je jich méně, a abychom měli dostatečnou volbu vhodných lexémů, museli jsme v našem výzkumu vytvořit širší prostřední pásmo.[8] K tomuto bodu se vrátíme později.

4. Metodologie

V rámci projektu „Acceptability and forced-choice judgements in the study of linguistic variation“ zkoumáme činitele ovlivňující rodilé mluvčí při hodnocení morfologických variant a při volbě vhodné varianty. Podle naší hypotézy existuje mezi korpusovou frekvencí a reakcemi rodilých mluvčích jistá souvislost (korelace) jak v hodnoceních variant, tak ve volbách variant. Navíc předpokládáme, že frekvence v korpusu bude řídícím činitelem ovlivňujícím rodilé mluvčí.

Na kolísání v Gsg a Lsg muž. rodu (vzor hrad, příklady typu toho jazyka/jazyku, *na hradě/na* hradu) již upozornila řada vědců v kvalitativních studiích syntaktické, stylistické a nářeční

[8] Totéž neplatilo např. pro Cvrčkovu gramatiku, protože tento systém se nesnaží o reálné rozdíly ve funkcích, má spíš za cíl popis frekvencí, a je proto vhodné, aby pásma byla pravidelněji rozložena.

variace.[9] V těchto pracích se upozorňuje na další činitele, včetně významu (v případě polysémních lexémů), syntaktického kontextu a regionálních rozdílů. Náš výzkum mířil jinam, tj. neopíral se o detailní zkoumání textů, nýbrž o četnost relevantních dokladů.

Dotázaní dostali k posouzení věty s oběma možnými tvary vybraných slov a k doplnění věty se slovy s vynechanými koncovkami (o hrad___). Abychom tuto souvislost změřili, museli jsme nejdříve operacionalizovat pojem *korpusová frekvence*; tj. zda je podstatná spíš absolutní frekvence nebo frekvence proporcionální. Vybrali jsme slova ve dvou pásmech absolutní frekvence a ve čtyřech pásmech proporcionální frekvence a tím jsme získali osm frekvenčních „buněk" na otestování:[10]

(5) Struktura *dotazníku podle použitých slov*

proporce {a/ě}, {a/u} / absolutní frekvence	**A**: 0–5%	**B**: 5–50 %	**C**: 50–95 %	**D**: 95–100 %
1: do 999 dokladů	**A1**	**B1**	**C1**	**D1**
Gsg	kožich, šuplík	obdélník, velín	čtvrtek, komín	oběd, ocet
Lsg	stadion, výraz	list, kanál	fotbal, strom	klášter, nos
2: 1000+ dokladů	**A2**	**B2**	**C2**	**D2**
Gsg	podzim,	sen	kout, rybník	kostel, národ
Lsg	zákoník	koncert, obvod	les, úřad	okres, stát
	pád, parlament			

Pro každou buňku jsme vybrali lexém, který jsme otestovali dvojím způsobem: hodnocení jednotlivých možných variant (např. *listě*–listu) a doplňování koncovek (např. list____) ve větách. Respondenti viděli každý lexém dvakrát v odlišných syntaktických kontextech charakteristických pro daný pád.

[9] Viz např. Bermel, 1993, 2004, 2010; Cummins 1995, Kasal, 1992; Klimeš, 1953; Kolařík, 1995; Rusínová, 1992; Sedláček, 1982; Štícha, 2009. K této problematice se věnuje širší diskuse i v mluvnicích, např. Petr a kol., 1986; Karlík – Nekula – Rusínová, 1995; Cvrček a kol., 2010.

[10] Původní rozložení počítalo s frekvenčními hranicemi 0–10 %, 10–50 %, 50-90 % a 95–100 %. Ale vzhledem k potřebě dostatečně velké volby lexémů v každé „buňce" včetně dvou pásem absolutní frekvence, jsme museli rozšířit pásma B a C až na pětiprocentní hranici (a i přesto jsme v buňce B2 našli jenom jedno vhodné slovo v Gsg). Data byla ověřena ve dvou velkých reprezentativních korpusech: SYN2005 a SYN2010. Vybraná slova musela spadat v obou korpusech do stejné buňky, co se týče absolutní a proporcionální frekvence, a mít minimální frekvenci v daném pádě nad sto dokladů.

Každý respondent odpovídal na dva typy dotazů: doplňování a hodnocení. Aby nedošlo k ovlivnění odpovědí pořadím úkolů, respondenti hodnotili a doplňovali různá slova: věty byly rozloženy do odlišných verzí dotazníků v tzv. „block design“ (uspořádání v blocích): ti, co hodnotili tvary v buňkách A1, B2, C1 a D2 doplňovali tvary z buněk A2, B1, C2 a D1 a naopak.[11] Toto uspořádání nám umožnilo zachovat přijatelnou délku dotazníků ve dvou paralelních verzích. Rozdílné verze zachycují část interakce mezi absolutními a relativními frekvencemi tvarů a vytváří možnost zkombinovat výsledky obou verzí (Cochran – Cox, 1957, s. 183-185). Zkombinované verze jsme potom zpracovali v komplexních analýzách rozptylu.

V dotazníkové akci provedené na vysokých školách, gymnáziích a pracovištích v různých částech České republiky jsme získali 587 vyplněných dotazníků.[12] Po vyřazení špatně vyplněných exemplářů a vyloučení odpovědí nerodilých mluvčích zbylo 551 použitelných dotazníků. Pomocí t-testů jsme srovnávali výsledky jednotlivých verzí dotazníků a zjistili, že pořadí otázek a úkolů nemělo na odpovědi respondentů významný vliv. Stejně tak se v našem vzorku nevyskytly významné rozdíly u proměnných jako věk, vzdělání či pohlaví.

5. Výsledky výzkumu přijatelnosti

Pomocí komplexních analýz rozptylu jsme chtěli zjistit, zda má proporcionální frekvence nebo absolutní frekvence větší dopad na to, jak respondenti hodnotili konkurující si tvary. Využili jsme k tomu statistický test analýzy rozptylu (ANOVA), která při velkém počtu respondentů (N=551) může být využita pro hodnocení na Likertově škále.[13] Výsledky ukázaly, že efekt proporcionální frekvence je ve všech případech významný a odpovídá za značnou část variace, ale pro absolutní frekvenci tomu tak nebylo.

V šesté tabulce jsou uvedeny výsledky pro proporcionální frekvenci v Gsg a v Lsg (každá paralelní verze je uvedena zvlášť). Důležité jsou především dvě hodnoty: p (pravděpodobnost, že zmíněný jev se tu vyskytl náhodně) a r (velikost efektu).

(6) *Proporcionální frekvence v* Gsg a Lsg*: výzkum přijatelnosti*

Gsg, verze 1: $F (1, 252) = 1305{,}97$, $p < 0{,}001$, $r = 0{,}92$

Gsg, verze 2: $F (1, 247) = 451{,}53$, $p < 0{,}001$, $r = 0{,}80$

Lsg, verze 1: $F (1, 253) = 489{,}89$, $p < 0{,}001$, $r = 0{,}81$

LSg, verze 2: $F (1, 251) = 223{,}97$, $p < 0{,}001$, $r = 0{,}69$

[11] Testovala se zároveň různá pořadí vět a pořadí úkolů: nejdříve doplňování, potom nejdříve hodnocení. Celou akci jsme zároveň opakovali s jinou sadou lexémů, abychom se pokud možno vyhnuli lexikálním efektům. O sestavení dotazníků viz např. Cowart (1997), Schütze (1996).

[12] V každé skupině respondentů bylo 16 mutací dotazníku rozdáno náhodně, aby se sociologický profil respondentů odrážel konzistentně ve všech mutacích.

[13] Tzv. Likertova škála se používá v dotaznících pro vyjádření souhlasu nebo náklonnosti stupňovaným způsobem (tj. jinak než „ano–ne“). V našem případě jde o sedmistupňovou škálu: 1 – *normální*, až 7 – *nepřípustné*.

Podle hodnoty p, které jsou konzistentně nižší než 0,05, lze usoudit, že výsledky jsou významné (to znamená, že se efekt frekvence pravděpodobně nevyskytuje náhodně). Hodnotou Cohenova r můžeme odhadnout velikost tohoto efektu: 0,1 je malý efekt, 0,3 je střední velikosti a 0,5 je velký efekt. Z toho vidíme, že efekty jsou ve všech případech velké, tj. tomuto jevu můžeme připsat velký podíl variace.

V sedmé tabulce jsou uvedeny výsledky pro absolutní frekvenci v Gsg a v Lsg:

(7) *Absolutní frekvence v* Gsg a Lsg*: výzkum přijatelnosti*

Gsg, verze 1: F (1, 252) = 106,66, p < 0,001, r = 0,55

Gsg, verze 2: F (1, 247) = 12,83, p < 0,001, r = 0,22

Lsg, verze 1: F (1, 253) = 16,55, p < 0,001, r = 0,25

LSg, verze 2: F (1, 251) = 223,97, p = 0,96

Hodnoty p ukazují, že výsledky jsou významné (tj. jsou nižší než 0,05) ve třech případech, nikoli ale pro druhou verzi Lsg. Hodnota Cohenova r spíše naznačuje, že jde ve dvou případech o menší efekt.

Proporcionální frekvence se zdá mít konzistentní, výrazný vliv na hodnocení uživatelů. Oproti tomu je vliv absolutní frekvence méně spolehlivý a méně výrazný.

Jeden možný důvod pro menší efekt absolutní frekvence může vyplývat ze šířky našich pásem (2 pásma oproti 4 pro proporcionální frekvenci). Rozhodli jsme se tedy počítat se skutečnými hodnotami absolutních frekvencí pro každé testovaný lexém, tj. bez použití pásem. Při absenci pásem nelze použít test ANOVA, ale je možné analyzovat data tzv. logistickou regresí[14]. Výsledky však byly ještě méně významné:

(8) Výsledky analýzy s přesnými hodnotami absolutní frekvence

	Absolutní frekvence	Abs. frekvence * Koncovka
Gsg, verze 1	F= 0,02, p = 0,881	F = 0,46, p = 0,50
Gsg, verze 2	F = 1,74, p = 0,19	F = 2,79, p = 0,95
Lsg, verze 1	F = 91,50, p = 0.99	**F = 72,43, p < 0,001**
LSg, verze 2	F = 7.97, p < 0,005	F = 0.28, p = 0,63

Zajímala nás významnost absolutní frekvence obecně (nezávisle na jejím spojení s koncovkou) a interakce mezi absolutní frekvencí a koncovkou. Zde posuzujeme významnost efektu hodnotou p (pravděpodobnost náhodnosti) v kombinaci s velikostí efektu, kterou

[14] V tomto případě šlo o zobecněný lineární smíšený model cílený na *zvolené skóre*, kde jsme mezi faktory přidali přesné absolutní frekvence daných lexémů.

odhadneme hodnotou F.[15] Výsledky dosáhly hranice významnosti a zároveň velikosti efektu jenom v jednom případě z osmi, který je označen tučně v tabulce 8. Nevýznamnost zbývajících výsledků naznačuje, že absolutní frekvence ovlivňuje hodnocení rodilých mluvčích rámcově, jak jsme viděli v tab. 7 (např. *časté – ne tak časté*), ale přesné absolutní frekvence nehrají při jejich rozhodování roli.

6. Výsledky výzkumu aktivního užití

Pro analýzu doplňování – kde měřené odpovědi nejsou hodnoty na škále, ale spíše volby z omezené řady ekvivalentních odpovědí (tj. koncovek) – jsme využili regresi. Šlo o zobecněný lineární smíšený model cílený na vybranou koncovku a mezi faktory jsme zadali *proporcionální* a *absolutní frekvence* daných lexémů.

Vypovídací schopnost (R^2) našeho modelu je pro všechny verze dotazníku vysoká.[16] Hodnota R^2 vychází z jednoduchého vzorce (viz níže) a vyjadřuje zhruba procentuální zlepšení, které model přináší nad jednoduchým modelem, ve kterém je vždy zvolena více frekventovaná koncovka, oproti „plnému" modelu, ve kterém jsou brány v úvahu všechny kombinace použitých faktorů:

$$R^2 = \frac{\text{Hodnota nového modelu} - \text{Hodnota výchozího modelu}}{\text{Hodnota plného modelu} - \text{Hodnota výchozího modelu}}$$

Pro naše čtyři sady slov jsme obdrželi následující hodnoty R^2: 76,4 % (Gsg 1), 81,0 % (Gsg 2), 64,1% (Lsg 1), 91,3 % (LSg 2), tj. ve všech případech jde o výrazné vylepšení modelu. To potvrzuje, že faktory, se kterými v našem modelu počítáme (např. absolutní a proporcionální frekvence) patří mezi důležité faktory při volbě koncovky.

Významnost našich proměnných měříme nadále hodnotou p a jejich relativní váhu odhadneme pomocí hodnoty F (v tabulce 9):

(9) *Proporcionální a absolutní* frekvence v Gsg a Lsg*: výzkum užití*

	Proporcionální frekvence	Absolutní frekvence
Gsg, verze 1	F= 157,52, p < 0,001	F = 81,10, p < 0,001
Gsg, verze 2	F = 122,62, p < 0,001	F = 21,66, p < 0,001
Lsg, verze 1	F = 90, 43, p < 0,001	F = 0,07, p = 0,80
LSg, verze 2	F = 91,50, p < 0,001	F = 1,99, p = 0,16

[15] Stručně řečeno, hodnota F se počítá z variace vysvětlené našim modelem dělené variací, kterou náš model nevysvětluje. Vyšší hodnoty F zpravidla indikují větší efekt.

[16] Tj. tím, že jsme do modelu zadali mj. tyto dva činitele, jsme o mnoho vylepšili předpověditelnost, kdy se která koncovka užívá.

Z tabulky 9 se dočteme, že proporcionální frekvence hraje vždy významnou roli (protože hodnota p je vždy nižší než 0,05). Váha této proměnné je větší než např. demografické charakteristiky respondentů, syntaktický kontext, apod. Absolutní frekvence oproti tomu hraje menší, ale významnou roli pouze v genitivu; v případě lokálu významná není.

7. Závěry

Naše předběžné statistické sondy do průzkumů ukazují, že proporcionální frekvence tvarů ve vyváženém korpusu, jako jsou korpusy SYN, je spolehlivě spojená s jejich přijatelností pro rodilé mluvčí a s frekvencí, s kterou rodilí mluvčí dané tvary vybírají. Absolutní frekvence tvarů ve vyváženém korpusu má mnohem méně spolehlivý účinek. Pokud jde o významný výsledek, dopad absolutní frekvence je ve všech případech méně výrazný než pro proporcionální frekvenci. Pro lokál má absolutní frekvence menší efekt: buď není významný vůbec, anebo je zanedbatelný oproti efektům jiných faktorů (např. proporcionální frekvence nebo syntaktického kontextu).

V této stati jsme upozornili na známou problematiku – na pojem frekvence – a k analýze našich dotazníkových dat jsme použili tři operacionalizace tohoto pojmu, abychom zjistili, která z nich tato data vysvětluje nejlépe: proporcionální frekvenci ve více kategoriích; absolutní frekvenci ve dvou kategoriích (vysoká/nízká); a absolutní frekvenci jako stupnici. Při formulaci jsme byli poněkud omezeni možnostmi našich dat. Pro jiné typy dat by bylo samozřejmě možné uvažovat i o jiných operacionalizacích tohoto pojmu. Volba jedné nebo druhé operacionalizace se zdá mít podstatný vliv na odpověď na naši obecnou otázku o souvislostech mezi *frekvencí* a chováním rodilých mluvčích.

LITERATURA

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: Z. Hladková, – P. Karlík (eds.), *Čeština – univerzália a specifika 5*, Praha: Nakladatelství Lidové Noviny, s. 163–171.

BERMEL, N. (2010): Variace a frekvence variant na příkladu tvrdých neživotných maskulin. In: S. Čmejrková – J. Hoffmannová – E. Havlová (eds.), *Užívání a prožívání jazyka,* Praha: Karolinum, s. 135–140.

BERMEL, N. – KNITTL, L. (2012a): Corpus frequency and acceptability judgments: A study of morphosyntactic variants in Czech. Corpus Linguistics and Linguistic Theory 8, s. 241–275.

BERMEL, N. – KNITTL, L. (2012b): Morphosyntactic variation and syntactic environments in Czech nominal declension: Corpus frequency and native-speaker judgments. Russian Linguistics, 36, s. 91–119.

BROWN, D. (2007): Peripheral functions and overdifferentiation: The Russian second locative. Russian Linguistics, 31, s. 61–76.

BYBEE, J. (2002): Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. Language Variation and Change, 14, s. 261–290.

BYBEE, J. (2006): From usage to grammar: The mind's response to repetition. Language, 82, s. 711–733.

BYBEE, J. (2007): Frequency of use and the organization of language. Oxford: Oxford University Press.

ČECH, R. (2012): Několik teoreticko-metodologických poznámek k Mluvnici současné češtiny. Slovo a slovesnost, 73, s. 208–216.

ČERMÁK, F. – KRÁLÍK, J. –KUČERA, K. (1997): Recepce současné češtiny a reprezentativnost korpusu (Výsledky a některé souvislosti jedné orientační sondy na pozadí budování Českého národního korpusu). Slovo a slovesnost, 58, s. 117–123.

COCHRAN, W. G. – COX, G. M. (1957): Experimental Designs (second edition). New York: John Wiley and Sons.

COWART, W. (1997): Experimental syntax: Applying objective methods to sentence judgments. Thousand Oaks, CA: Sage Publishers.

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.

CVRČEK, V. – KODYTEK, V. (2013): Ke klasifikaci morfologických variant. Slovo a slovesnost, 74, 139–145.

CVRČEK, V. A KOL. (2010): *Mluvnice současné češtiny*. Praha: Karolinum.

ČESKÝ NÁRODNÍ KORPUS – SYN2005, SYN2010. Ústav Českého národního korpusu FF UK, Praha 2005, 2010. Dostupný z WWW: http://www.korpus.cz

DIVJAK, D. (2008): On (in)frequency and (un)acceptability. In: B. Lewandowska-Tomaszczyk (ed.), Corpus linguistics, computer tools and applications – State of the art, Frankfurt: Peter Lang, s. 213–233.

HALLIDAY., M. A. K. (1992): Language as system and language as instance: The corpus as a theoretical construct. In: J. Svartvik (ed.), Directions in Corpus Linguistics, Berlin: Mouton de Gruyter, s. 61–77.

KARLÍK, P. – NEKULA, M. – RUSÍNOVÁ, Z. A KOL. (1995): *Příruční mluvnice češtiny.* Praha: Nakladatelství Lidové Noviny.

KASAL, J. (1992): Dublety a jejich užití. Philologica, 65, 107–114.

KLIMEŠ, L. (1953): Lokál singuláru a plurálu vzoru „hrad“ a „město“. *Naše řeč*, 36, s. 212–219.

KOLAŘÍK, J. (1995): Dynamika ve flexi substantiv běžně mluveného jazyka ve Zlíně. In: D. Davidová, (ed.), K diferenci*aci současného mluveného jazyka.* Ostrava: Universitas Ostraviensis, Facultas Philosophica, s. 79–83.

KRÁLÍK, J. – ŠULC, M. (2005): The representativeness of Czech corpora. International Journal of Corpus Linguistics, 10, s. 357–366.

PETR, J. A KOL. (1986): Mluvnice *češtiny*. Praha: Academia.

RUSÍNOVÁ, Z. (1992): Některé aspekty distribuce alomorfů (genitiv a lokál sg. maskulin). *Sborník prací filozofické fakulty brněnské univerzity,* A 40, s. 23–31.

SCHÜTZE, C. (1996): The empirical base of linguistics: Grammaticality judgments and linguistic methodology. Chicago: University of Chicago Press.

SEDLÁČEK, M. (1982): V Záhřebě i v Záhřebu. *Naše řeč,* 65, s. 11–15.

ŠTÍCHA, F. (2009): Lokál singuláru tvrdých neživotných maskulin (ve vlaku vs. v potoce): úzus a gramatičnost. Slovo a slovesnost, 70, s. 193–220.