Morfologická produktivita v diachronní perspektivě: příklad sufixů *-mento/-zione* ve staré italštině od 13. do 16. století

Pavel Štichauer

ABSTRACT:
Morphological productivity in the diachronic perspective: the case of suffixes *-mento/-zione* in Old Italian from the 13th to the 16th century. This paper deals with morphological productivity in diachrony, in particular it addresses the issue of the quantitative evaluation of productivity within a given time span. Adopting Baayen's (1992; 2001; 2008) corpus-based quantitative approach which considers productivity as the probability of encountering a new type when sampling a large corpus, the paper shows the evolution of two competing suffixes *-mento/-zione* in Old Italian from the 13th to the 16th Centuries. On the basis of four separate corpora drawn from LIZ 4.0 (*Letteratura Italiana Zanichelli*), it is demonstrated how the productivity of the suffix *-mento*, within the time span of four centuries, remains constant, while the suffix *-zione* displays diachronic variability. Apart from diachronic considerations regarding this situation, the paper also highlights some technical aspects, such as the use of LNRE models (implemented in the package *zipfR*, a tool for lexical statistics in *R*, cf. Baroni — Evert, 2006; Evert — Baroni, 2007; Baayen, 2008), as well as some well-known limitations and constraints inherent in quantitative analyses of diachronic corpora.

KLÍČOVÁ SLOVA / KEY WORDS:
diachronní korpusy, lexikální statistika, produktivita, sufixy *-mento/-zione*, stará italština 13.– 16. století, program *zipfR*
diachronic corpora, lexical statistics, productivity, suffixes *-mento/-zione*, Old Italian 13th–16th Cent., package *zipfR*

1. ÚVOD

Cílem tohoto článku[1] je ukázat možnosti a meze diachronní aplikace kvantitativního pojetí morfologické produktivity, které vychází zejména z prací R. Haralda Baayena (1992, 2001, 2008, 2009). Jak bude patrné z dalšího výkladu, Baayen pojímá produktivitu v kvantitativním slova smyslu jako míru pravděpodobnosti, s níž budeme v dostatečně velkém korpusu nacházet nová slova, která bychom mohli považovat za neologismy, svědčící právě o produktivním slovotvorném prostředku. Určení takové pravděpodobnosti a zejména její srovnání napříč různými korpusy — obsahujícími texty pocházející z různých časových intervalů — je bezpochyby zajímavý úkol.

1 Tento článek vznikl za podpory projektu Univerzity Karlovy Progres 4, Q10, Jazyk v proměnách času, místa, kultury. Rád bych vyjádřil svůj dík Ondřeji Tichému za pročtení první verze textu. Děkuji rovněž dvěma anonymním recenzentům za mnoho zásadních připomínek, které jsem se ve finální verzi pokusil co nejvíce zohlednit.

V tomto textu se pokusím takové srovnání nabídnout pro dvojici italských deverbálních sufixů *-mento/-zione*, a to napříč čtyřmi (sub)korpusy pokrývajícími čtyři tradičně definovaná století (13., 14., 15. a 16. století).

Volbu sufixů i časového intervalu se teď pokusím stručně zdůvodnit. Sufixy *-mento/-zione* patří v současné italštině mezi základní přípony, jimiž se derivují *nomina actionis* (např. *rimodernare — rimodernamento* „modernizovat — modernizace“, *banalizzare — banalizzazione* „banalizovat — banalizace“), a v některých případech představují konkurenční prostředky bez zjevného významového rozdílu (např. *congelazione — congelamento* „zmražení, zamražení“). Oba sufixy patří rovněž mezi určité diachronní konstanty, lišící se právě jen proměnlivou produktivitou mezi starou a současnou italštinou; zároveň nejsou přímo závislé na určitém žánru, na rozdíl např. od sufixu *-anza*, který lze v hojném množství nalézt v básnickém jazyce 13. století (srov. Coletti, 1993, s. 7; Castellani, 2000, s. 503).

Hovoříme-li o „staré italštině“, je potřeba upozornit na určitou pojmovou neustálenost: lze takto označovat pouze florentštinu velkých trecentistů (Dante, Boccaccio, Petrarca), popř. šířeji koncipovanou literární „toskánštinu“, již počátkem 16. století kodifikoval Pietro Bembo jako ideální literární jazyk a jejíž textový kánon poměrně přesně vymezil. Vzhledem k chronologickému intervalu, který nás zde bude zajímat, budu pochopitelně pojímat starou italštinu jako literární jazyk, který kromě zmíněných velikánů italské literatury bude obsahovat rovněž další texty 13. století (jako je např. *Novellino*) a také texty, které nepatří jen do umělecké prózy či poezie, ale také do odborných žánrů (historické spisy, filosofické traktáty apod.). Tam, kde to bude pro otázky produktivity zásadní, upozorním na možné dopady, které zařazení či naopak vyřazení některých textů mohou provázet.

2. DEFINICE PRODUKTIVITY

Pojem produktivita patří bezesporu k tomu, co lze s G. Dalovou (2003, s. 3) nazvat „společnou terminologickou výbavou“ každého lingvisty. Aby však bylo zřejmé, v jakém smyslu zde budu tohoto pojmu používat, lze vyjít z definice, s níž přišla D. Corbinová (1987, s. 177):

> *„… produktivita ve skutečnosti označuje zároveň pravidelnost výsledných formací,* ***dostupnost*** *daného afixu, tzn. možnost tvoření nedoložených slov a možnost vyplňovat mezery doloženého lexika, a* ***výnosnost****, tj. aplikovatelnost na velký počet bází a/nebo produkce velkého počtu doložených formací.“*[2]

Pojmy *dostupnosti* a *výnosnosti*, v anglické terminologii *availability* a *profitability*, které pro francouzské *disponibilité* a *rentabilité* navrhl Carstairs-McCarthy (1992, s. 37) a pře-

2 *„… la productivité désigne en fait à la fois la régularité des produits de la règle, la disponibilité de l'affixe, c'est-à-dire précisément la possibilité de construire des dérivés non attestés, de combler les lacunes du lexique attesté, et la rentabilité, c'est-à-dire la possibilité de s'appliquer à un grand nombre de bases et/ou de produire un grand nombre de dérivés attestés.“*

vzal pak Bauer (2001, s. 205–209; 2008, s. 322), užitečně zachycují dva klíčové aspekty běžně chápané produktivity, totiž aspekt ryze kvalitativní, kdy hovoříme o strukturní přítomnosti či nepřítomnosti určitého slovotvorného prostředku, a aspekt kvantitativní, kdy již dostupný prostředek zkoumáme z hlediska *počtu* výsledných formací, které díky němu vznikají. Oba tyto aspekty jsou samozřejmě do určité míry neoddělitelné, neboť „výnosný" prostředek bude nutně „dostupný", různá míra výnosnosti pak může signalizovat některá podstatná omezení, kterými je takto dostupný prostředek vázán. Kvalitativní studium produktivity je tedy především výzkumem celé řady restrikcí (fonologických, morfologických, syntaktických či sémantických), zatímco cílem kvantitativního výzkumu bude stanovit počet výsledných formací, popř. právě stanovit míru pravděpodobnosti, s níž ke vzniku nových slov bude docházet.

3. KVANTITATIVNÍ POJETÍ PRODUKTIVITY: *DICTIONARY-BASED* A *CORPUS-BASED*

Ponecháme-li kvalitativní výzkum stranou, neboť není předmětem tohoto článku, lze odlišit dva základní postupy kvantitativně pojaté analýzy: na straně jedné tzv. *dictionary-based*, na straně druhé *corpus-based*. Ve slovníkově založeném výzkumu jsme omezeni na výzkum jediné veličiny, totiž na počet *lemmat* (či *typů*).[3] Přesto lze i v tomto případě diachronní aspekt produktivity zachytit, máme-li k dispozici data ci prvního doložení daného slova.[4] Tento přístup tedy měří *type frequency*, tj. počet lemmat. Naproti tomu korpusově založený výzkum pracuje nejen s *type frequency* (V), ale také s počtem *výskytů/tokenů* (N), a to vzhledem k velikosti korpusu (F). Právě vztah mezi (V) a (N) stojí v základu Baayenova pojetí.

4. BAAYENOVA KONCEPCE: PRODUKTIVITA JAKO *PRAVDĚPODOBNOST*

Baayenovo pojetí je v zásadě velmi prosté. Vztah mezi *token frequency* (N) a *type frequency* (V) může být nahlédnut jako funkce V(N): počet V je funkcí N, tj. se vzrůstající hodnotou N poroste i hodnota V; N je dáno velikostí korpusu.[5] Tento vztah lze zachytit pomocí „křivky nárůstu slovníku" (*vocabulary growth curve*).

Jako příklad můžeme použít již zpracovanou sadu dat, italská slovesa s iterativním prefixem *ri-*, na které Baroni & Evert (2006) ukazují možnosti lexikální sta-

3 Mezi termíny *lemma/type* zde nebudu činit žádný rozdíl a budu pracovat s běžnou distinkcí *type/token* (srov. např. Baroni, 2009), popř. *lemma/výskyt*.

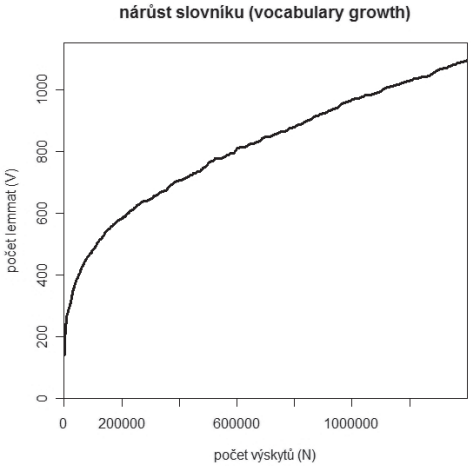
4 Např. Plag (1999) používá *OED* k výzkumu sloves tvořených sufixy *-ize* a *-ify*, omezuje se přitom na 20. stol. Bauer (2001) ukazuje diachronní aspekt substantiv derivovaných pomocí přípon *-ation* a *-ment*, a to ve větším časovém úseku (od 16. do 20. stol.). Oba lingvisté zároveň dobře popisují výhody i nevýhody slovníkově založeného výzkumu (srov. zejména Plag, 1999, s. 96–100; též Bauer, 2001, s. 156–161).

5 Později pak bude nutné odlišit (N) jako počet tokenů omezený na daný vzorek relevantních slov a již zavedený (F), tj. velikost korpusu.

tistiky v programu *R* (a za použití balíčku *zipfR*, který sami vytvořili).[6] Zpracovaný frekvenční seznam těchto prefigovaných sloves pochází z korpusu *La Repubblica* (F = 330 mil. tokenů) a obsahuje 1098 lemmat s celkovou frekvencí 1 399 898 tokenů (tj. celková frekvence všech těchto sloves s prefixem *ri-*). Postupnou excerpci, při níž si pro stanovený interval (např. každých 200 tis. tokenů) zapíšeme hodnoty (V) a (V_1), tj. počet všech lemmat a počet všech lemmat, která se v korpusu vyskytla zatím jen jednou (*hapax legomena*), zachycuje tabulka č. 1, empirickou[7] křivku nárůstu slovníku pak obrázek č. 1

N	V	V_1
200 000	583	176
400 000	707	213
600 000	810	259
800 000	881	274
1 000 000	969	306
1 200 000	1029	314

TABULKA 1: Počet lemmat (V) a *hapax legomena* (V_1) sloves s prefixem *ri-* na základě dat z korpusu *La Repubblica* (Baroni & Evert, 2006).



OBRÁZEK 1: Empirická křivka nárůstu italských sloves s prefixem *ri-* na základě dat z korpusu *La Repubblica* (Baroni & Evert, 2006).

Baayen zároveň definuje i *tempo*, jakým slovník narůstá (*vocabulary growth rate*). Rychlost nárůstu lze určit ze *sklonu* křivky v daném bodě, který se vyjádří poměrem

6 Viz níže, kde je *zipfR* blíže představen. Všechny grafy jsou vytvořeny v programu *R* (https://www.r-project.org/), R Core Team (2017), s instalovaným balíčkem *zipfR*.

7 Jak uvidíme později, empirická křivka (na rozdíl od interpolované/extrapolované) zachycuje reálně naměřené hodnoty.

všech *hapax legomena*, tj. lemmat, která se objevují v korpusu jen jednou, a celkovým počtem jejich výskytů, tedy $P = V_1 / N$ (viz tabulka č. 2).

slovotvorný prostředek	počet lemmat (V)	výskyty (N)	počet *hapax legomena* (V_1)	P (V_1/N)
ri + sloveso	1098	1 399 898	346	0,00024

TABULKA 2: Hodnoty V, V_1, N a míry pravděpodobnosti P(V_1/N) sloves s prefixem *ri-* z korpusu *La Repubblica* (Baroni & Evert, 2006).

Hodnota P zde vyjadřuje pravděpodobnost nalezení nového lemmatu uvnitř korpusu; počet typů, které se vyskytují jen jednou, je vydělen celkovým počtem všech ostatních výskytů sledovaných typů s prefixem *ri-*: *„The growth rate is a probability, the probability that, after having read N tokens, the next token sampled represents an unseen type, a word type that did not occur among the preceding N tokens."* (Baayen, 2008, s. 223). Baayen tedy navrhuje vidět produktivitu — ve smyslu oné výnosnosti — jako pravděpodobnost, že narazíme na nové slovo, když budeme postupně procházet daný korpus (popř. to lze interpretovat jako nějakou časovou dimenzi). Srovnání hodnot P pro určité slovotvorné procesy s sebou nese také určité obtíže, zejména nutnost disponovat srovnatelnými korpusy a zajistit vzorky o stejných velikostech N (viz k tomu Gaeta & Ricca, 2002, 2003, 2006; Štichauer 2009b, s. 34–51, 2009c). Tyto obtíže lze do určité míry překonat použitím teoretických modelů, které vzápětí uvedu.

Dříve než se dostanu k diachronním aplikacím, je třeba poukázat na celou řadu studií, které tuto metodu aplikují na synchronní data, např. Baayen & Renouf, 1996; Gaeta & Ricca, 2002, 2003, 2006; Baroni, 2007; Dal, 2003; Dal et al., 2007.

5. DIACHRONNÍ APLIKACE

Diachronní aplikace tohoto postupu, ačkoli ji komplikují určité problémy, o kterých budu v závěru hovořit, spočívá v porovnání hodnot P (nebo v prostém vizuálním porovnání růstových křivek) v daných časových intervalech t_1, t_2, … t_n (srov. např. Dalton-Puffer & Cowie, 2002; Lüdeling & Evert, 2005; Scherer, 2007; Säily & Suomela, 2009; Štichauer, 2009a, b; 2015a, b; Rácz, Papp & Hay, 2016, Sect. 6). Klíčovou otázku, jak jednotlivé časové úseky stanovit či vybrat, zde musím jen stručně nastínit. Na jedné straně je možné převzít tradiční periodizaci (např. na století) a dále ji členit podle určitých kritérií (např. interval jedné generace, půlstoletí apod.); na straně druhé je možné (a metodologicky zajímavější) periodizaci „nechat vyčíst" přímo z dat na základě kvantitativně signifikantních rozdílů; v takovém případě získáme periodizaci, v níž jsou patrné body zlomu, tj. etapy, ve kterých došlo k výrazné kvalitativní změně. Takové etapy pak mohou být časově velmi různorodé (od několika století po desetiletí) (srov. k tomu zejména Gries & Hilpert, 2008, 2012; Hilpert & Gries, 2009). V následující aplikaci se však přidržím tradiční periodizace do století, která odpovídá běžně vymezeným obdobím italského jazyka a jako celek (od 13. do 16. století)

pak vytváří určitou homogenní periodu, kterou Tesi (2001, s. 5–9) charakterizuje jako období psané italštiny s implicitní normou založenou na kanonických textech (více k volbě této periodizace srov. Štichauer 2009b, s. 70–71).

5.1. CHARAKTERISTIKA KORPUSŮ

Data, s nimiž budu pracovat, pocházejí z CD-ROMu textů italské literatury LIZ 4.0 (2001). Jde o kolekci textů, jež sice nesplňuje charakteristiky vyváženého diachronního korpusu, ale která může dobře posloužit k vytvoření takových korpusů. LIZ 4.0 obsahuje 1000 textů italské literatury od 13. do počátku 20. století a zahrnuje téměř všechna zásadní díla nejen literárního jazyka, ale i celou řadu významných regionálních tradic. CD-ROM ovšem umožňuje selekce textů podle různých kritérií, např. období, žánru, autora apod. Periodizace je již pevně nastavena, takže je možné vytvářet subkorpusy pouze v řádech století (Duecento, Trecento…), popř. vytvářet subkorpusy na základě selekce jednotlivých textů/autorů. Pro účely tohoto výzkumu jsem takto vytvořil čtyři subkorpusy odpovídající zmíněným čtyřem stoletím. Po eliminaci textů, které rozhodně do kánonu „staré italštiny" zařadit nelze,[8] vykazují subkorpusy charakteristiky shrnuté v tabulce č. 3.

Období / subkorpus	Velikost v tokenech	Počet textů
13. století	732 114	40
14. století	3 565 818	55
15. století	2 675 016	46
16. století	10 604 452	231

TABULKA 3: Základní parametry subkorpusů 13.–16. století vytvořených z LIZ 4.0.

Jak je patrné, jde o subkorpusy, které jsou nesrovnatelné nejen kvantitativně, ale rovněž kvalitativně, neboť každé období obsahuje různé procento zastoupených žánrů; někde je tato situace nutně dána skladbou dochovaných textů (např. pro 13. století), někde je naopak dána tím, které texty LIZ 4.0 obsahuje (např. pro 16. století by bylo možné nejen některé texty eliminovat, ale naopak i některé přidat, ne všechny jsou v LIZ 4.0 zastoupeny) (k problematické povaze historických korpusů srov. např. Claridge, 2008). Přesto je tento vzorek se všemi omezeními, která z něho plynou, určitým východiskem pro alespoň hrubé odhady vývoje produktivity, o které nám půjde.

5.2. FREKVENČNÍ SEZNAMY A JEJICH ZPRACOVÁNÍ

Pro každé století (=subkorpus) je východiskem *frekvenční seznam* všech lemmat zakončených na sufix *-mento* a *-zione*. Abych získal co nejúplnější seznam — byť po-

8 Pro 13. století je to např. Bonvesin de la Riva, jehož texty jsou psány ve staré milánštině (popř. v tom, co bude později označováno jako *lombardská koiné*); pro 15. století je to např. extravagantní text *Hypnerotomachia Poliphili* Francesca Colonny (viz níže).

chopitelně velmi hrubý, obsahující mnoho „extrakčního šumu“ —, pracoval jsem v první fázi s prostým regulárním výrazem *.*ment/o/i*, respektive *.*zion/e/i*, který se ihned kvůli ortografickému rozptylu ukázal jako nedostatečný; grafické varianty se týkají zejména sufixu *-zione*, který v doložených textech v závislosti na kritické edici může vystupovat v podobě *-sione, -tione, -ttzione* atd., proto jsem postupně dotaz rozšiřoval.

Ve druhé fázi jsem přistoupil ke zcela manuální lemmatizaci; tato fáze představovala náročný krok, neboť shrnutí některých variant pod jedno lemma nebylo vždy jednoznačné. Např. u dvojice *acendimento — accendimento* („rozsvícení“) je zcela zřejmé, že jde o jedno lemma lišící se pouze ortograficky, ale u dvojice *adimandamento — domandamento* („otázka, dotaz, dotazování“) už zdaleka tak jasno není. Základním kritériem — zejména proto, že jde o studium produktivity, tedy o analýzu vztahu mezi bázovým slovesem a jeho derivátem — byla právě vazba na sloveso. V tomto případě máme tedy dvě lemmata, neboť v textech je doložena dvojice *adimandare — adimandamento*, stejně tak jako *domandare — domandamento* („ptát se“ — „otázka“).

Ve třetí fázi bylo nutné z takto lemmatizovaných seznamů odstranit všechna slova, která nelze označit za deriváty s danými sufixy. Gaeta a Ricca (2002, 2003, 2006) navrhli — pro synchronní účely — sérii pěti kritérií, která lze (s určitými modifikacemi) aplikovat i na diachronní data, jak lze na základě následujících příkladů dobře vidět.

1) Silná „neprůhlednost“, slabý derivační vztah k bázovému slovesu — např. *impressione / imprimere* („dojem“ vs. „vtisknout“).[9]
2) Tzv. *baseless formations*, formace bez jakékoli verbální báze, většinou zřejmé latinismy — např. *accezione* (vs. *accettazione* od *accettare* „přijmout“), z lat. *acceptio*.
3) Nominální báze — např. *vasellamento* („nádobí“), obvykle jde o formace s kolektivním významem.
4) Tzv. *inner derivational cycles*, tj. případy, kdy sufix *-mento/-zione* není tím posledním připojeným; derivát je bází k další derivaci, jako např. *indeterminazione — *indeterminare* (ale *determinazione — indeterminazione*).
5) Výpůjčky — např. *entergezione* (lat. *interiectio*), *saramento* (fr. *serment*).

Je samozřejmě možné uvažovat o určité redukci těchto skupin, např. formace bez jakékoli slovesné báze by mohly figurovat ve skupině č. 5 jako výpůjčky, ale pro praktické účely — mj. proto, aby byla povaha každé formace správně a detailně posouzena — je daleko výhodnější pracovat s touto sérií mírně se překrývajících kritérií (např. proto, že tak lze odlišit reálné výpůjčky v daném diachronním okamžiku, jako např. *saramento* z fr. *serment*, od nespočtu formací, které přecházejí přímo z latiny bez

9 Zde je dobře vidět, že manuální lemmatizace je nutná, neboť jen na základě kontextu lze odlišit, kdy je *impressione* již lexikalizovaným substantivem ve významu „dojem“ a kdy je naopak přímočarým derivátem, který zachovává význam bázového slovesa, tj. *imprimere — impressione* „tisknout — tištění“.

svého bázového slovesa a které by jako „výpůjčky" pak představovaly dobrou třetinu či polovinu všech formací se sufixem *-zione*).[10]

Z výsledných frekvenčních seznamů pak extrahujeme pouze seznam frekvencí, který pak v prostém formátu txt (kde je na prvním řádku *f* a pak již konkrétní frekvence) zpracuje *zipfR* v programu *R*.[11]

5.3. PŘEHLED ZPRACOVANÝCH DAT

Výše uvedené „pročišťovací" kroky nám pak nabízejí následující definitivní data pro jednotlivé subkorpusy, jak je shrnuje tabulka č. 4.

Období / subkorpus	Sufix	V	V_1	N
13. století	*-mento*	280	137	2093
	-zione	143	47	1653
14. století	*-mento*	455	206	6475
	-zione	398	119	7717
15. století	*-mento*	351	172	3457
	-zione	548	177	6679
16. století	*-mento*	583	281	14026
	-zione	722	194	32031

TABULKA 4: Počty lemmat (V), *hapax legomena* (V_1), tokenů (N) pro sufixy *–mento/-zione* od 13. po 16. století.

Jak je patrné z této tabulky, jednotlivé soubory jsou do veliké míry nesrovnatelné. Mezi 13. stoletím, tj. nejmenším subkorpusem, a 16. stoletím je velký rozptyl. Přesto lze alespoň částečně vysledovat určité tendence. 13. a 14. století je charakteristické tím, že stále dominuje sufix *-mento*; formace se sufixem *-zione* však představují slova, jejichž *token frequency* postupně rychle narůstá, zatímco z hlediska počtu typů nedosahují úrovně konkurenčního sufixu *-mento*. Situace se začíná proměňovat od 15. století, tj. s výrazným vlivem humanistické latiny, a vrcholí pak v 16. století. Přesto nelze jednoznačně říci, jak vzápětí uvidíme, že by od 16. století začal sufix *-zione* nabývat na produktivitě. Jak je z tabulky patrné, počty *hapax legomena* jsou daleko nižší, narůstá naopak *token frequency* (a dosahuje dvojnásobku N konkurenčního *-mento*). Abychom

10 Slova se sufixem *-zione* jsou v tomto ohledu velmi problematická. Situace do 16. století se výrazně odlišuje od následného vývoje, kdy tento sufix nabývá nezpochybnitelnou autonomii, zatímco ještě do 16. století bychom mohli v zásadě hovořit o masivním procesu slovotvorné výpůjčky (srov. Štichauer 2015a, 2015b).

11 Do příkazového řádku v *R* pak píšeme *název.tfl <- read.tfl(„název.txt")*; po zadání *summary (název.tfl)* získáme *zipfR object for frequency spectrum*, s nímž pak pracujeme při tvorbě teoretických modelů (viz níže). Na oficiálních stránkách programu *zipfR* (http://zipfr.r-forge.r-project.org/) je k dispozici ke stažení nejen celý balíček, ale i různé pomocné skripty, jakož i ukázkové datové sady.

mohli všechny čtyři vzorky porovnat, nezbývá než vytvořit teoretické modely se srovnatelnými parametry, tj. zejména s identickým počtem N.

6. TEORETICKÉ MODELY A JEJICH INTERPRETACE

Takové srovnatelné modely lze vytvořit na základě rozdělení frekvencí slov známého jako LNRE (*Large Numbers of Rare Events*) (srov. Baayen 2001; 2008, s. 222–236, kap. 6.5. *Models for lexical richness*). Balíček *zipfR* obsahuje tři dosud dostupné modely, GIGP (*Generalized Inverse Gauss-Poisson*; srov. Baayen, 2001, s. 89–93), ZM a fZM (*Zipf-Mandelbrot / finite Zipf-Mandelbrot*; srov. Evert, 2004).[12] Aby mohl program *zipfR* tyto modely vytvořit, vychází z *frekvenčního spektra*, jež získá na základě již zmíněného frekvenčního seznamu. Frekvenční spektrum (srov. Baroni, 2009) zobrazuje počty typů *V* s danou frekvencí *V(m)* od *m* = 1 (*hapax legomena*), *m* = 2, *m* = 3 až po *m* = *nejvyšší frekvence* doloženého typu. Toto spektrum pak slouží jako vstupní data pro zmíněné tři modely, které produkují teoretické hodnoty (*expected values*) E(V), E(V_1) pro jakoukoli hodnotu N. Buď lze pracovat s empirickou hodnotou N daného vzorku (tedy např. pro 13. století u sufixu *-mento* s N = 2093) a nepřekračovat ji (tím získáme prostřednictvím binomické interpolace predikované hodnoty pro jakoukoli hodnotu N až do N = 2093), anebo lze tuto hranici překročit a použít *extrapolaci*, při níž získáváme predikované hodnoty pro vyšší N, než je reálná velikost vzorku.[13]

Vzhledem k tomu, že mezi našimi čtyřmi korpusy je obrovský rozptyl (nejmenší N = 1653, největší N = 32031), je třeba najít nějakou společnou hodnotu N, při níž jednotlivé sady můžeme koherentně porovnat. Jako příklad volíme (víceméně extrémní) hodnotu N = 20000 a model ZM, který se po sérii testů extrapolační kvality ukázal jako nejspolehlivější.[14] Predikované hodnoty E(V), E(V_1), tj. hodnoty, které produkuje teoretický model ZM pro danou velikost vzorku N = 20 000, včetně P = V_1 / N zachycuje tabulka č. 5.[15]

Jak je patrné nejen z predikovaných hodnot, ale i z prostého náhledu nárůstových křivek (kde je na ose *y* zobrazena predikovaná hodnota V v závislosti na N, tj. E[V(N)]), zůstává sufix *-mento* naprosto konstantní ve své produktivitě zejména

12 K matematickým aspektům všech tří modelů, ke kterým zde nemohu kompetentně poskytnout dostatečné résumé, srov. Baayen 2001, 2008. Vynikající popis všech vlastností LNRE modelů a jejich východisek lze také nalézt v prezentacích z kurzu Stefana Everta a Marca Baroniho z r. 2006 (dostupné z WWW: http://zipfr.r-forge.r-project.org/esslli2006.html). Konkrétní výsledky jednotlivých modelů na zde uváděných diachronních datech srov. Štichauer, 2009b, s. 74–78.

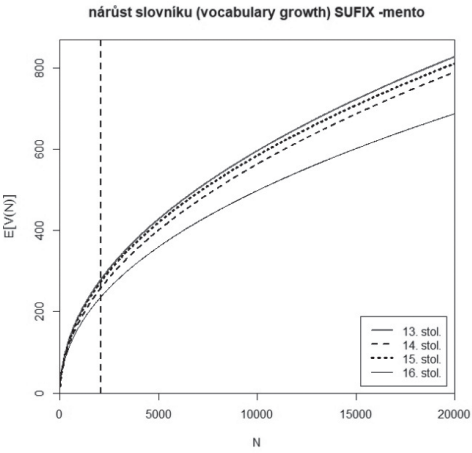
13 K interpolaci a extrapolaci srov. Baayen, 2001, s. 63–76; 2008, s. 232–236 .

14 K otázce této spolehlivosti, zejména ke *goodness-of-fit* a k intervalům spolehlivosti srov. Evert & Baroni, 2005; Baayen, 2008, s. 233–234; Štichauer, 2009a. Ke konkrétním výsledkům na zkoumaných diachronních datech srov. Štichauer, 2009b, s. 46–50, 74–78.

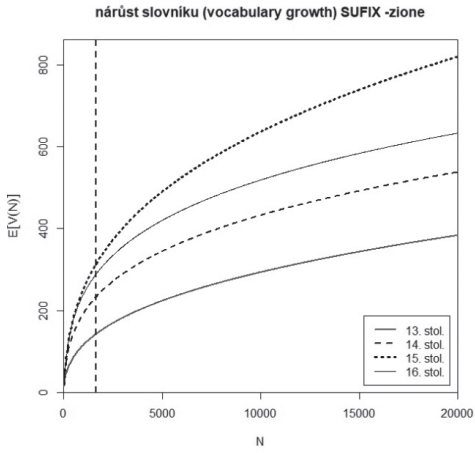
15 Hodnoty jsou zaokrouhleny.

Sufix	N	E(V)	$E(V_1)$	$P(V_1/N)$
-mento (13. stol.)	20 000	827	390	0,020
-mento (14. stol.)	20 000	789	383	0,019
-mento (15. stol.)	20 000	811	382	0,019
-mento (16. stol.)	20 000	687	318	0,016
-zione (13. stol.)	20 000	384	147	0,007
-zione (14. stol.)	20 000	538	167	0,008
-zione (15. stol.)	20 000	821	295	0,015
-zione (16. stol.)	20 000	634	178	0,009

TABULKA 5: Predikované hodnoty E(V), $E(V_1)$ a míra produktivity P při identickém N = 20000 pro sufixy *-mento/-zione* na základě modelu ZM.



OBRÁZEK 2: Sufix *-mento* od 13. do 16. století při N = 20000; model ZM.



OBRÁZEK 3: Sufix *-zione* od 13. do 16. století při N = 20000; model ZM.

první tři sledovaná století; teprve 16. století pak přináší mírný pokles.[16] Tento výsledek je důležitý, neboť potvrzuje některé běžné teze o vývoji tohoto sufixu. Na jedné straně lze doložit, že některé formace se sufixem *-mento* z úzu postupně mizí a jsou nahrazeny jednak formacemi se sufixem *-zione*, jednak deverbálními jmény s nulovým sufixem (či jmény vzniklými konverzí). Např. velmi časté *consolamento* je od 16. století postupně nahrazováno formací *consolazione* („útěcha"). Tuto tendenci lze velmi dobře doložit na mnoha dalších slovech (zejména pak od 16. století). Hůře doložitelnou tendencí — zejména proto, že deverbální substantiva s nulovým sufixem je obtížné automaticky vyhledat — je pak přechod nebo přímá derivace (či konverze, dle teorie) substantiv jako *abbandono* („opuštění"), *abbraccio* („obejmutí"), kterým ve starších fázích konkurují formace *abbandonamento, abbracciamento*.

16 Srov. Štichauer, 2009a, s. 145, k intervalům spolehlivosti, z nichž je patrné, že mezi 13., 14. a 15. stol. skutečně není signifikantní rozdíl, zatímco 16. století se již významně liší.

Podstatně odlišná situace pak panuje u sufixu *-zione*. Vizuální srovnání křivek nám nabízí poměrně jasný obrázek o proměnlivé situaci; přesto je analýza predikovaných hodnot v tab. 5 důležitá, neboť ukazuje, že tempo nárůstu zůstává pro tři století v zásadě shodné (13., 14. a 16.), zatímco jen 15. století vykazuje hodnoty srovnatelné s konkurenčním sufixem *-mento*. Situace se však výrazně nemění ani po 16. století; minimálně do 19. století je *-zione* ve slabší produktivitě v zásadě konstantní. Teprve později, mezi 19. a 20. století se stává zcela autonomním prostředkem, a to především díky nárůstu verbálních bází se sufixem *-ificare* a *-izzare*, které jsou dnes hlavním „inputem" pro derivaci deverbálních jmen se sufixem *-zione* (srov. k tomu Štichauer 2015a, 2015b, kap. 5).

Jak je tedy zřejmé, tvorba srovnatelných teoretických modelů může bezpochyby posloužit k základní kvantifikaci produktivity jednotlivých procesů. Interpretace zjištěných rozdílů či naprosté neměnnosti je pak otázkou ryze lingvistickou: někde je ve hře externí tlak (silný vliv latiny v období humanismu), někde pak také interní změny slovotvorných možností (jako je právě komplementární nárůst produktivity verbálních sufixů *-ificare*/*-izzare*).

7. ZÁVĚRY

Studií, jež by aplikovaly baayenovský korpusově založený přístup na diachronii, stále není mnoho (srov. Rácz, Papp & Hay, 2016, s. 699). Důvody jsou zřejmé. Předně jde o problematické složení diachronních korpusů. Ve hře je nejen jejich heterogenní, nevyvážená povaha,[17] ale i problematická lemmatizace, kterou komplikuje slabá standardizace starších textů (např. ortografický rozptyl), a především malý rozsah, který spolehlivost zde popsaných metod, vyžadujících velké objemy dat, výrazně snižuje. Všechny tyto nedostatky se projevují víceméně známým způsobem.

Tzv. *clustering effects* (Evert, 2005) jsou v diachronních korpusech — alespoň tedy v těch, s kterými jsme pracovali v této studii — daleko výraznější. Jde o typické shlukování frekvencí slov v určitých textech. Např. v subkorpusu 14. století lze nalézt 246 výskytů slova *consolazione*, z nichž 60 jich pochází z *Dopisů Kateřiny Sienské*; ještě výraznější — a v podstatě fatální — frekvenční vychýlení představuje v subkorpusu 16. stol. slovo *raccomandazione* („doporučení"): z celkového počtu 144 tokenů jich dobrých 138 nalezneme v jediném textu (*Dopisy* Torquata Tassa).

Specifickým problémem je pak povaha *hapax legomena*. Ideálně vzato bychom měli pokud možno zaručit, že doložená *hapax legomena* představují v daném úseku

17 Vyvážené diachronní korpusy existují, ale tato vyváženost jde na úkor jejich velikosti. Např. korpus MIDIA (http://www.corpusmidia.unito.it/), realizovaný na univerzitě v Turíně, má celkovou velikost asi 7,5 mil. tokenů (srov. výše uvedený korpus pro 16. stol. o velikosti zhruba 10 mil. tokenů), je složen z 800 textů; korpus je rozdělen do 5 diachronních úseků (subkorpusů), každý z nich je tvořen 25 texty o přesně vymezené velikosti 8000 tokenů pro každý žánr. Dále pak samozřejmě existují žánrově vymezené diachronní korpusy, které sice nejsou veliké, ale jsou více homogenní (např. *Corpus of Early English Correspondence* (CEEC) či korpus italské korespondence 19. století *Corpus Epistole Ottocentesco Digitale* (CEOD)).

neologismy, tj. zcela nově utvořená slova dokládající produktivitu slovotvorného prostředku.[18] Jestliže je tento požadavek v zásadě nesplnitelný i pro synchronní korpusy (srov. k tomu Dal, 2003), o to horší je situace v případě diachronních korpusů, a to ze dvou důvodů. Za prvé proto, že vzhledem k malé velikosti korpusů budou mezi *hapax legomena* i slova, která pak v rozsáhlejších korpusech dosahují vyšších frekvencí. Za druhé proto, že i když např. bezpečně víme, že dané slovo vytvořil jako první autor daného textu, je zcela běžné, že ono slovo pak v tomtéž textu použije vícekrát. V tomto smyslu nám takový neologismus nešťastně vypadne ze série *hapax legomena*, které k výpočtu produktivity použijeme. S tím souvisí i někdy víceméně umělá, artistní povaha takových neologismů, která svědčí spíše o individuální „kreativitě" než o spontánní „produktivitě" (srov. k tomu Plag, 1999, s. 13–16). Tento problém je zcela zřetelný např. v 15. stol. u již zmíněného specifického textu Francesca Colonny *Hypnerotomachia Poliphili*, jenž je charakteristický přítomností mnoha efemérních kreací jak se sufixem *-mento*, tak se sufixem *-zione*.[19]

Přes všechna tato omezení vede diachronní aplikace kvantitativně založeného přístupu k morfologické produktivitě k zajímavým výsledkům, které mohou upřesnit či vyvrátit některé běžně přijímané teze o vývoji slovotvorných procesů. Diachronní výzkum slovotvorby tak získává užitečný nástroj, který bude v budoucnu jistě ještě využit.[20]

LITERATURA

Baayen, R. H., & Renouf, A. (1996). Chronicling the *Times*. Productive Lexical Innovations in an English Newspaper. *Language, 72*, 69–96.

Baayen, R. H. (1992). Quantitative aspects of morphological productivity. In: G. Booij & J. van Marle (Eds.), *Yearbook of Morphology 1991* (s. 109–149). Dordrecht: Kluwer.

Baayen, R. H. (2001). *Word Frequency Distributions*. Dordrecht: Kluwer.

Baayen, R. H. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Baayen, R. H. (2009). Corpus linguistics in morphology: Morphological productivity. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook. Volume 2* (s. 899–919). Berlin: Mouton de Gruyter.

Baroni, M., & Evert, S. (2006). The *zipfR* package for lexical statistics: A tutorial introduction, version 2014. Dostupné z WWW: <http://zipfr.r-forge.r-project.org/materials/zipfr-tutorial.pdf>.

Baroni, M. (2007). I sensi di ri-. Un'indagine preliminare. In R. Maschi, N. Penello & P. Rizzolatti (Eds.), *Miscellanea di studi linguistici offerti a Laura Vanelli* (s. 163–171). Udine: Forum.

Baroni, M. (2009). Distributions in text. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook*.

18 Zde je tedy neologismus chápán široce a zahrnuje jak prokazatelně nově derivované slovo, jež si např. vynucuje syntaktický kontext, tak také různé efemérní okazionalismy, autorská slova apod. Srov. k těmto definicím např. Janovec, 2013, s. 106; Martincová, 2016.

19 Štichauer 2009b, s. 104–116, nabízí dva modely pro 15. stol., jeden pro celý korpus, druhý bez zmíněného textu, včetně podrobné diskuse některých formací, jakož i kvantitativních rozdílů obou modelů.

20 Srov. např. Naccarato, 2016.

Volume 2, (s. 803–822). Berlin: Mouton de Gruyter.
Bauer, L. (2001). *Morphological Productivity*. Cambridge: Cambridge University Press.
Bauer, L. (2008). Productivity: theories. In P. Štekauer & R. Lieber (Eds.), *Handbook of Word-Formation* (s. 313–332). Dordrecht: Kluwer Academic Publishers.
Carstairs-McCarthy, A. (1992). *Current Morphology*. London: Routledge.
Castellani, A. (2000). *Grammatica storica della lingua italiana. 1. Introduzione.* Bologna: il Mulino.
Claridge, C. (2008). Historical Corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook. Volume 1* (s. 242–259). Berlin: Mouton de Gruyter.
Coletti, V. (1993). *Storia dell'italiano letterario. Dalle origini al Novecento*. Torino: Einaudi.
Corbin, D. (1987) : *Morphologie dérivationnelle et structuration du lexique*, 2 voll., Tübingen: Niemeyer.
Dal, G., Fradin, B., Grabar, N., Lignon, S., Namer, F., Plancq, C., Yvon, F., & Zweigenbaum, P. (2007). Linguistic prerequisites to the calculation of morphological productivity and first results. Talk presented at *Journées ATALA*, Paris, November 10, 2007.
Dal, G. (2003). Productivité morphologique: définitions et notions connexes. *Langue française*, *140*, 3–23.
Dalton-Puffer, C., & Cowie, C. (2002). Diachronic word-formation and studying changes in productivity over time: Theoretical and methodological considerations. In J. E. Díaz Vera (Ed.), *A Changing World of Words. Studies in Englih Historical Lexicography, Lexicology and Semantics* (s. 410–437). Amsterdam: Rodopi.
Evert, S., & Baroni, M. (2005). Testing the extrapolation quality of word frequency models. *Proceedings of Corpus Linguistics 2005*, Dostupné z WWW: <http://www.corpus.bham.ac.uk/PCLC/>.
Evert, S., & Baroni, M. (2007). zipfR: Word frequency distributions in R. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Sessions*, 29–32.
Evert, S. (2004). A simple LNRE model for random character sequences. *Proceedings of JADT 2004*, 411–422.
Evert, S. (2005). *The Statistics of Word Cooccurrences: Word Pairs and Collocations* (nepublikovaná disertační práce). Stuttgart: University of Stuttgart.
Gaeta, L., & Ricca, D. (2002). Corpora testuali e produttività morfologica: i nomi d'azione in due annate della *Stampa*. In R. Bauer & H. Goebl (a cura di), *Parallela IX. Testo — variazione — informatica. Text — Variation — Informatik* (s. 223–249). Wilhelmsfeld: Gottfried Egert Verlag.
Gaeta, L., & Ricca, D. (2003). Frequency and productivity in Italian derivation: A comparison between corpus-based and lexicographical data. *Italian Journal of Linguistics / Rivista di Linguistica*, *15*(1), 63–98.
Gaeta, L., & Ricca, D. (2006). Productivity in Italian word formation: A variable-corpus approach. *Linguistics*, *44*(1), 57–89.
Gries, S. Th., & Hilpert, M. (2008). The identification of stages in diachronic data: variability-based neighbour clustering. *Corpora*, *3*(1), 59–81.
Gries, S. Th., & Hilpert, M. (2012). Variability-based neighbor clustering: a bottom-up approach to periodization in historical linguistics. In T. Nevalainen & E. C. Traugott (Eds.), *The Oxford Handbook of the History of English* (s. 134–144). Oxford: Oxford University Press.
Hilpert, M., & Gries, S. Th. (2009). Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, *24*(4), s. 385–401.
Janovec, L. (2013). Neologie. In M. Martinková & O. Uličný (Eds.), *Studie k moderní mluvnici češtiny 4. Dynamika českého lexika a lexikologie* (s. 105–130). Olomouc: Univerzita Palackého v Olomouci.
LIZ 4.0 (2001). *Letteratura italiana Zanichelli. CD-ROM dei testi della letteratura italiana*,

a cura di Pasquale Stoppelli ed Eugenio Picchi. Bologna: Zanichelli.

Lüdeling, A., & Evert, S. (2005). The Emergence of Non-Medical *-itis*. Corpus Evidence and Qualitative Analysis. In S. Kepser & M. Reis (Eds.), *Linguistic Evidence. Empirical, Theoretical, and Computational Perspectives* (s. 315–333). Berlin: Mouton de Gruyter.

Martincová, O. (2016). Neologismus. In P. Karlík, M. Nekula & J. Pleskalová (Eds.), *Nový encyklopedický slovník češtiny* (s. 1163–1164). Praha: NLN.

Naccarato, C. (2016). A corpus-based quantitative approach to the study of morphological productivity in diachrony: The case of *samo*-compounds in Russian. In H. Christ, D. Klenovšak, L. Sönning & V. Werner (Eds.), *A Blend of MaLT. Selected Contributions from the Methods and Linguistic Theories Symposium 2015* (s. 133–153). Bamberg: University of Bamberg Press.

Plag, I. (1999). *Morphological Productivity. Structural Constraints in English Derivation.* Berlin: Mouton de Gruyter.

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rácz, P., Papp, V., & Hay, J. (2016). *Frequency and Corpora*. In A. Hippisley & G. Stump (Eds.), *The Cambridge Handbook of Morphology* (s. 685–704). Cambridge: Cambridge University Press.

Säily, T., & Suomela, J. (2009). Comparing type counts: The case of women, men and *-ity* in early English letters. In A. Renouf & A. Kehoe (Eds.), *Corpus Linguistics: Refinements and Reassessments* (s. 87–109). Amsterdam: Rodopi.

Scherer, C. (2007). The role of productivity in word-formation change. In J. C. Salmons & S. Dubenion-Smith (Eds.), *Historical Linguistics 2005: Selected Papers from the 17th International Conference on Historical Linguistics, Madison, Wisconsin, 31 July — 5 August 2005* (s. 257–271). Amsterdam: John Benjamins.

Štichauer, P. (2009a). Morphological productivity in diachrony: the case of the deverbal nouns in *-mento*, *-zione* and *-gione* in Old Italian from the 13th to the 16th century. In F. Montermini, G. Boyé & J. Tseng (Eds.), *Selected Proceedings of the 6th Décembrettes* (s. 138–147). Somerville, MA: Cascadilla Proceedings Project.

Štichauer, P. (2009b). *La produttività morfologica in diacronia: i suffissi* -mento, -zione *e* -gione *in italiano antico dal Duecento al Cinquecento*. Praha: Karolinum.

Štichauer, P. (2009c). Approccio quantitativo alla produttività morfologica: alcuni sviluppi recenti. *Écho des études romanes*, *V*(1–2), 7–25.

Štichauer, P. (2015a). From emergent availability to full profitability: The diachronic development of the Italian suffix *-zione* from the 16th to the 20th century. In: S. Augendre, G. Couasnon-Torlois, D. Lebon, C. Michard, G. Boyé & F. Montermini (Eds.) *Proceedings of the Décembrettes. 8th International Conference on Morphology* (s. 319–326). Toulouse: CNRS & Université Toulouse.

Štichauer, P. (2015b). *La formazione delle parole in diacronia. Studi di morfologia derivazionale dell'italiano tra il Cinquecento e l'Ottocento*. Praha: Karolinum.

Tesi, R. (2001). *Storia dell'italiano. La formazione della lingua comune dalle origini al Rinascimento*. Roma-Bari: Laterza.

Pavel Štichauer | Ústav románských studií FF UK
<pavel.stichauer@ff.cuni.cz>