Kvantitativní analýza ortografické variability v *Korpusu rané anglické korespondence*[1]

Ondřej Tichý

ABSTRACT:

A quantitative analysis of spelling variation in the *Corpus of Early English Correspondence* (CEEC). The paper explores trends in spelling variation as reflected in Early English correspondence (15th–17th c.) on the material of the *Corpus of Early English Correspondence* (CEEC).

Overall change in spelling variation has so far been commented on only in relatively general terms and never on quantitative grounds. There is, of course, no doubt about the general direction of the change (towards greater standardization, though not in a straightforward manner) and its basic characteristics, such as its slower pace in private documents compared to the spelling of professional publications, but the data to support the assertions as well as precise definitions of spelling variation or regularisation have not yet been, to our knowledge, provided.

This paper introduces a novel methodology for the quantification of spelling variation and regularity, which allows a more objective assessment of its change and which also makes use of the metadata provided by the *CEEC*: such as gender, letter authenticity or relationship/kinship between the author and the recipient.

The paper explores interactions of such variables from the diachronic perspective using quantified levels of spelling regularity. The measure introduced for this purpose is based on weighted information (Shannon) entropy, as a measure of predictability of spellings of individual functionally defined types, and its calculation is partly based on the morphological tagging of the parsed version of the *Corpus*.

KLÍČOVÁ SLOVA / KEY WORDS:
korespondence, korpus, entropie, variace, ortografie
correspondence, corpus, entropy, variation, spelling

ÚVOD

Dějiny anglické ortografie jsou popsány především jako dějiny utváření standardů a dějiny postupné standardizace. Dosavadní práce se soustředily především na proces utváření: z jakých zdrojů čerpal, kdo byli jeho nositelé, jakým způsobem se rozšiřoval, na jeho systémový popis a na popis změn tohoto systému.

Není přitom v zásadě sporu o tom, že proces standardizace ortografie vede ke snižování ortografické variability a že v dějinách angličtiny, alespoň od pozdní střední angličtiny, se rozvíjející se standard postupně šíří, a kolísavost anglického pravopisu tedy i postupně klesá. Zřejmě proto, že tento trend je celkem očividný, dočkal se v odborné literatuře zatím spíše obecných komentářů, např.:

1 Tato práce vznikla za podpory projektu Kreativita a adaptabilita jako předpoklad úspěchu Evropy v propojeném světě, reg. č.: CZ.02.1.01/0.0/0.0/16_019/0000734, financovaného z Evropského fondu pro regionální rozvoj.

„as the fifteenth century progressed so a universal stabilised orthography … was increasingly widely used" (Scragg, 46), „after 1550 we find a … greater stability and regularity of spelling in private documents", "up to the final fixing of spelling circa 1650" (*ibid.*, 68), případně „the 15th century saw a steady movement towards a fixed spelling that very much resembles the spelling of Modern English" (Upward & Davidson, 2011, s. 174).

Stávající práce si klade za cíl formulovat metodologii pro kvantifikaci ortografické variability a tradičně obecně popsaný trend potvrdit, respektive konfrontovat s korpusovými daty na úžeji vymezeném vzorku textů dopisů z let 1410–1680. Jak jsme uvedli výše, byť trend samotný je zpochybnitelný jen stěží, kvantitativní metodologie umožní přesněji sledovat jeho výkyvy, porovnávat jeho vývoj mezi různými jazykovými varietami (privátní/neprivátní texty, nářečí, žánry) a v neposlední řadě pomůže lépe definovat, co se konkrétně ortografickou variabilitou myslí.

METODOLOGIE

Vymezení variability může být jednoduché: K variabilitě dochází vždy, když „[existuje] více forem pro jeden význam, n. více významů pro jednu formu" (Cvrček, 2017), o ortografické variabilitě pak mluvíme, když pro jednu funkci užíváme více než jednu grafickou formu.[2] Tuto definici je však třeba dále upřesnit: především které funkce máme na mysli?

Za funkční budeme pro účely této práce považovat jen rozlišení v rámci jedné variety, a ne rozdíly např. mezi dialekty. Jinými slovy, *varietní* tvary (ať již odrážející různou výslovnost, nebo různé pravopisné standardy) stejného slova považujeme za funkčně nerozlišené, a tedy *variantní*. Např. (britské) <colour> a (americké) <color> považujeme za kolísání pravopisu, protože v rámci jedné variety nejde o funkční rozlišení.

Nezohledňujeme také ortografické rozlišení vyšších funkčních rovin, např. pragmatické nebo stylové. Takové rozlišení je v angličtině v rámci lexikálně-gramatických jednotek řídké, a pokud se vyskytuje, jde především o rozlišení pomocí archaických morfologických tvarů (např. kontrast slovesných tvarů 2. os. př. času <sayest> a <say>). V případě změn funkčních kategorií v průběhu zkoumaného období, které vedly ke vzniku archaických tvarů (např. zde zmíněné formální ztrátě druhé osoby j. č. u přítomných slovesných tvarů), pak předpokládáme funkční rozlišení na úrovni flektivní gramatické morfologie v celém sledovaném období. Naopak formy rozlišující funkce sémantické nebo gramatické za funkčně odlišné považujeme.[3]

2 Psanou formu jazyka považujeme v tomto smyslu za nezávislou na formě mluvené a nedefinujeme tak variabilitu ortografie v závislosti na mluvených formách. Vzhledem k použitému materiálu (ohledně spolehlivosti materiálu viz níže) také nerozlišujeme variabilitu grafickou a grafémickou, byť nám jde především o tu druhou.

3 Jde nám přitom ale pouze o rovinu konkrétních slov, výskytů, např. kolísání forem <dared> a <durst>, nikoliv o rovinu systémovou, nezabýváme se tedy obecně např. varia-

Takto vymezenou variabilitu můžeme kvantifikovat, např. velmi jednoduše vydělením počtu distinktních grafických tvarů (dále formy) počtem lexikálně-gramatických jednotek (dále typy[4]), tedy jako v literatuře populární „counting of spellings" (např. Crystal, s. 394). V korpusové lingvistice ale takový postup těžko obstojí, mimo jiné zcela opomíjí četnost jednotlivých forem (dále tokeny).

Jiným způsobem, jak na základě těchto veličin míru variability typů kvantifikovat, je kalkulovat podíl počtu forem na počet jejich výskytů (dále tokeny), takže čím více forem na počet tokenů by se v textu objevilo, tím vyšší míru variability by text vykazoval. Pokud ale blíže prozkoumáme konkrétní data, zjistíme, že takto jednoduše pojatá kvantifikace často neodpovídá tomu, jak míru variability intuitivně chápeme. Jako příklad uveďme formy slova OTHER a minulého času slovesa SHOULD spolu s počty tokenů těchto forem ve dvou po sobě jdoucích dekádách (viz tabulku 1). V případě slova OTHER poměr mezi první a druhou dekádou šestnáctého století výrazně klesá z 0,093 na 0,067, intuitivně bychom ale variabilitu druhé dekády označili zřejmě za vyšší. Naopak v případě minulých tvarů slovesa SHOULD poměr mezi padesátými a šedesátými lety stejného století výrazně stoupá, intuitivně však formální kolísání spíše klesá. Tento způsob kvantifikace totiž pomíjí rozložení tokenů a forem. Jak ho zohlednit?

OTHER	**1500–09**	**1510–19**	**SHOULD-pret**	**1550–59**	**1560–69**
nodder	1	—	*shold*	8	7
nother	1	6	*sholde*	1	4
oder	—	3	*should*	5	1
odyr	—	2	*shoulde*	—	1
other	39	57	*shuld*	10	1
othre	2	6	*shulde*	5	—
celk. tokenů	43	74		29	14
forem/tokenů	0,093	0,067		0,172	0,357

TABULKA 1: Formy a tokeny slova OTHER a minulého času slovesa SHOULD.

Domníváme se, že intuitivně by míra variability nebo kolísání měla být tím vyšší, čím menší šanci máme, že odhadneme formu výskytu konkrétního typu v textu — neboli je tím vyšší, s čím nižší pravděpodobností ji budeme schopni určit/odhadnout. Tato definice sice může naznačovat, že podíl forem a jejich tokenů je správnou mírou variability, protože jde o vyjádření pravděpodobnosti, nám však nejde o pravděpodobnost výskytu jedné konkrétní formy, ale o rozložení pravděpodobností (resp. o míru nejistoty ohledně) forem konkrétních typů v textu.

Míru nejistoty systému je možné vyjádřit tzv. informační nebo také Shannonovou entropií, přičemž entropii jednoho konkrétního typu definujeme jako „kde *n* je

bilitou nesamostatných morfémů typu variantních koncovek préterita <-ed> vs. <-t(e)> apod.

4 Jak z uvedeného vyplývá, zde užívané lexikálně-gramatické jednotky zcela neodpovídají v korpusové lingvistice běžně užívanému termínu *typ*, viz seznam užitých typů v příloze 1.

počet forem daného typu, *f* je frekvence formy (počet tokenů formy) a *t* frekvence typu (počet tokenů všech forem daného typu)". Ve výše zmíněných příkladech budou hodnoty entropie pro OTHER 0,176 (první dekáda) a 0,363 (druhá dekáda), zatímco pro SHOULD 0,627 (padesátá léta) a 0,552 (šedesátá léta). Tato čísla ukazují opačný a intuitivnější trend než prostý poměr forem a typů.

Pokud chceme použít entropii jako míru variability v textu, v němž se vyskytuje více typů, z nichž každý vykazuje vlastní míru entropie, nemůžeme entropie jednotlivých typů pouze jednoduše sčítat. Jednotlivé typy jsou v textu zpravidla zastoupeny různě často, měly by tak k míře variability přispívat různou měrou (např. entropie 0,58 za 213 tokenů 11 forem spojky IF by neměla mít stejnou váhu v celkové entropii textu jako entropie 0,93 minulého času slovesa SHALL s 5582 tokeny celkem v 61 formách). Lze namítat, že počet tokenů je již započítán v entropii daného typu. To je sice pravda, představme si ale text (viz tabulku 2), který je tvořen 100 tokeny pouze dvou typů — jeden z nich má pouze jednu formu, a tedy nulovou entropii, vyskytuje se však v textu v 98 tokenech, druhý má dvě formy, z nichž každá je v textu reprezentována jedním tokenem, a vykazuje tedy entropii zhruba 0,3. Pokud entropie pouze sečteme, bude výsledná entropie textu opět zhruba 0,3. Snížíme-li však v textu počet tokenů prvního typu na 50 a zároveň zvýšíme počet tokenů obou forem druhého typu na 25 za každou, získáme intuitivně podstatně variabilnější text, který by ale po prostém sečtení entropií obou typů vykazoval entropii stále shodnou — zhruba 0,3. Z tohoto důvodu nejprve entropie jednotlivých typů násobíme jejich relativní frekvencí a až následně sčítáme, tím v uvedeném případě získáme váženou entropii prvního textu zhruba 0,006, zatímco vážená entropie druhého textu bude řádově vyšší, přibližně 0,15.

		tokenů formy	**entropie typu**	**tokenů formy**	**entropie typu**
		text 1	**text 1**	**text 2**	**text 2**
typ 1	forma 1.1	98	0	50	0
typ 2	forma 2.1	1	0,3	25	0,3
	forma 2.2	1		25	
součet entropie			0,3		0,3
vážená entropie			0,006		0,15

TABULKA 2: Rozdílná agregace entropie.

Míru variability textu tedy počítáme jako relativní frekvencí vážený součet entropií jednotlivých typů. K takto navržené míře se sluší dodat dvě důležité charakteristiky. Zaprvé platí to, že čím vyšší je její index, tím vyšší je měřená variabilita, přičemž hodnota 0 znamená nulovou variabilitu. Zadruhé kvůli vážení dílčích entropií jednotlivých typů jejich relativní frekvencí není výsledná míra v pravém smyslu mírou entropie, ale považujeme ji prostě za míru variability.

KORPUS A ZPRACOVÁNÍ DAT

Problém s vyhledáváním variantních tvarů dle navržené metodologie v historických korpusech spočívá obvykle především v chybějícím značkování. Dostupné historické korpusy angličtiny nejsou lemmatizovány, natož aby měly značkované jednotlivé morfologické kategorie a rozlišovaly tak lexikálně-gramatické jednotky definované výše. Jak tedy v korpusu takové jednotky identifikovat a vyhledat zároveň všechny jejich tvary? Ruční vyhledávání, případně značkování celého korpusu nebo jedné dostatečně velké části by vyžadovalo značnou časovou investici, rozhodli jsme se proto pracovat s omezeným, ale specifickým vzorkem dat.

Řada helsinských historických korpusů (Rissanen et al.) existuje v syntakticky a morfologicky značkovaných verzích, které sice nejsou lemmatizovány ani nemají komplexně značkovány morfologické kategorie, mají ale značkovány alespoň slovní druhy. Výhodou použitého morfologického značkování této řady korpusů je, že ač jde primárně o značkování slovních druhů (POS tagging), některé lexikální jednotky (především jde o slova gramatická nebo o slova z uzavřených slovních druhů, např. OTHER jako zájmeno či spojka) jsou značeny samostatně, a v některých případech jsou dokonce značkovány jejich vybrané gramatické kategorie (např. pro sloveso DO je značeno zvlášť přítomné participium, pasivní participium, infinitiv, minulé tvary, imperativ, minulé participium a prézentní tvary, pro ostatní jednotky je ale morfologické značkování mnohem méně důkladné). Pragmaticky jsme tedy vybrali vzorek dat omezený na nejdetailněji značkované slovní druhy, resp. lexikální jednotky — konkrétně na slovesa BE, DO, HAVE, modální slovesa, předložku/spojku FOR, existenciální THERE, spojky WHETHER a IF a příslovce OTHER a SUCH. Všechny výskyty odpovídající těmto značkám jsme ze značkované verze korpusu *CEEC* (*PCEEC*) spolu se všemi metadaty vyexportovali pomocí online korpusového manažeru ČNK *Kontext*.

Celkem jsme tímto způsobem získali 238 659 tokenů v 930 formách spolu s příslušnými metainformacemi. Ty bylo třeba ručně roztřídit do výše definovaných funkčních typů, resp. je dodatečně roztřídit v případech, kdy značkování samotného korpusu nezachycuje celou morfologickou pestrost angličtiny 15.–17. století. Celkem jsme všechny výskyty roztřídili do 79 typů, např. modální slovesa, značkovaná všechna dohromady jako „MD", jsme ručně dodatečně roztřídili na 26 typů, konkrétně na přítomné tvary, přítomné tvary 2. a 3. os. sg., tvary préterita, tvary 2. os. sg. prét. sloves CAN, DARE, MAY, MUST, NEED, OWE, SHALL a WILL (ne všechny typy byly využity pro všechna slovesa).[5]

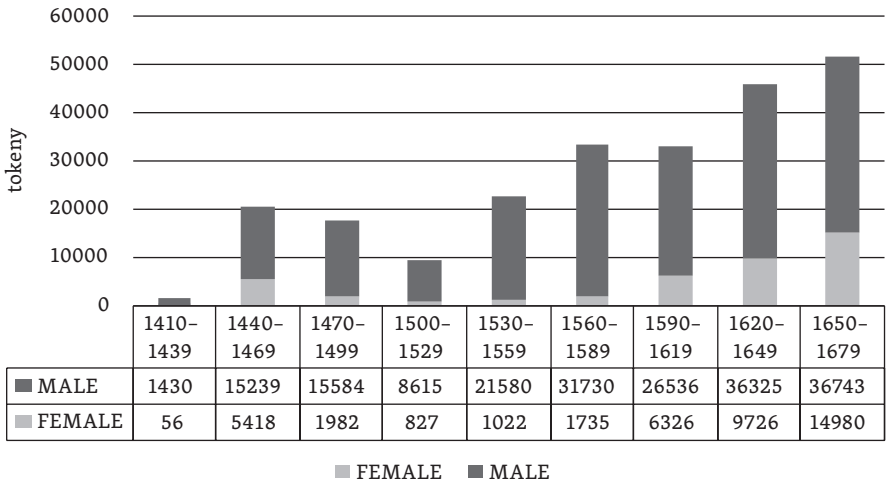
Jak jsme uvedli výše, korpus *CEEC* sestává z dopisů a díky tomu je také doplněn metadaty, která ke každému textu v korpusu (dopisu) doplňují řadu sociolingvistických proměnných. Pro následující analýzu jsme se kromě doby sepsání dopisu rozhodli využít také metadat o pohlaví autora, autenticitě dopisu (autograf, zapsaný jinou osobou) a příbuzenském vztahu autora a příjemce dopisu. Pro účely našeho výzkumu jsme kategorie metadat zjednodušili způsobem naznačeným v tabulce 3.

5 Přehled všech typů je uveden v příloze 1.

kód	původní kategorie	naše kategorie
A (A/C)	autograf	autograf (autograph)
B	autograf málo známé osoby	
C (C/A)	kopie nebo dopis psaný písařem	písařský (secretarial)
D (D/A)	autenticita neznámá	neznámá (unknown)
FN	nejužší rodina	rodina (family)
FO	širší rodina	
FS	rodinný sloužící	
TC	přátelé	ostatní (others)
T	ostatní	

TABULKA 3: Zjednodušení kategorií metadat CEEC.

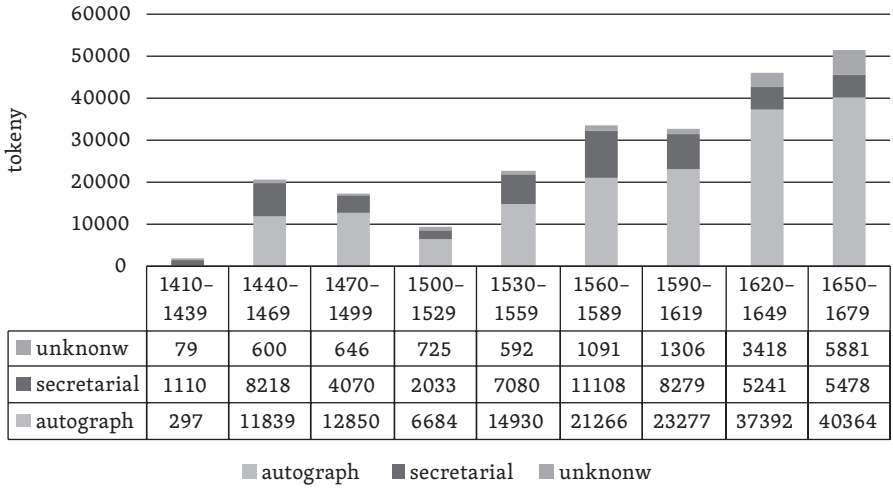
Ne všechny kategorie jsou v korpusu pochopitelně zastoupeny stejnou měrou. Pro naši analýzu je však důležité vzít rozdílné zastoupení metadat naznačené v následujících grafech v úvahu.



	1410–1439	1440–1469	1470–1499	1500–1529	1530–1559	1560–1589	1590–1619	1620–1649	1650–1679
MALE	1430	15239	15584	8615	21580	31730	26536	36325	36743
FEMALE	56	5418	1982	827	1022	1735	6326	9726	14980

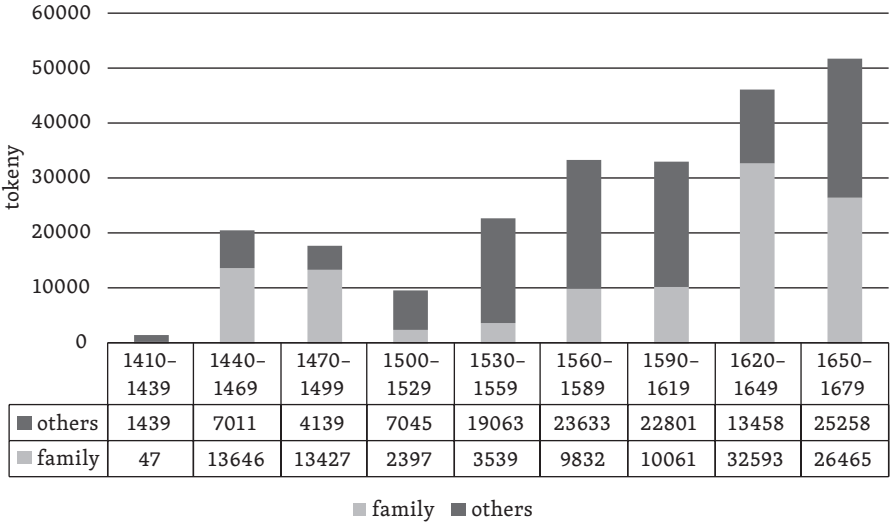
GRAF 1: Rozložení tokenů excerpovaných z dopisů psaných ženami (female) a muži (male).

Jak je patrné z grafu 1, zastoupení tokenů z dopisů psaných ženami, zjednodušeně řečeno zastoupení žen, je poměrně nízké, a to především v prvním, ale i v několika prostředních obdobích. I z důvodu takového rozložení jsme se rozhodli data analyzovat nikoliv po obvyklejších dekádách, kde by rozsah dat zvláště v sociolingvisticky úzce definovaných kategoriích byl příliš malý, ani po *Helsinským korpusem* zavedených obdobích, která by data rozdělovala jen na čtyři neúplná období, ale kompromisně po třicetiletých úsecích, které mohou odpovídat např. generačním cyklům.



	1410–1439	1440–1469	1470–1499	1500–1529	1530–1559	1560–1589	1590–1619	1620–1649	1650–1679
unknonw	79	600	646	725	592	1091	1306	3418	5881
secretarial	1110	8218	4070	2033	7080	11108	8279	5241	5478
autograph	297	11839	12850	6684	14930	21266	23277	37392	40364

GRAF 2: Rozložení tokenů v dopisech psaných samotnými autory (autograph) nebo písaři (secretarial).



	1410–1439	1440–1469	1470–1499	1500–1529	1530–1559	1560–1589	1590–1619	1620–1649	1650–1679
others	1439	7011	4139	7045	19063	23633	22801	13458	25258
family	47	13646	13427	2397	3539	9832	10061	32593	26465

GRAF 3: Rozložení tokenů v dopisech příbuzným (family) a ostatním adresátům (others).

Jak je patrné z grafů 2 a 3, v průběhu času se zvyšuje zastoupení autografů na úkor dopisů psaných písaři, zatímco zastoupení rodinných příslušníků mezi adresáty kolísá.

Samotný výpočet variability za jednotlivé časové úseky a sociolingvistické kategorie byl proveden v jazyce R. Jelikož rozsah zpracovaných dat se mezi časovými úseky výrazně odlišoval (zvlášť s ohledem na některé sociolingvistické kategorie a jejich kombinace), přistoupili jsme v zájmu vyšší výpovědní hodnoty našich výsledků ke

značení konfidenčních intervalů[6] založených na statistické metodě *bootstrapping*. Tato metoda umožňuje odhadnout přesnost pozorování pomocí opakovaného náhodného vzorkování dat. Konkrétně jsme k výpočtům konfidenčních intervalů využili balíček *boot* (funkce boot a boot.ci) parametrizovaný na tisíc náhodných vzorků, metodu basic a rozsah intervalu 95 % (viz Canty, 2017; Davison, 1997).

V části věnované materiálu je třeba zmínit také problematičnost využití některých v korpusu obsažených dat pro analýzu ortografie. Zcela odhlédnuto od problému autorství, respektive písařského autorství, je nutné vzít v potaz i spolehlivost moderního přepisu, který nejprve proběhl z rukopisu do edic a poté z edic do korpusu. Autoři korpusu se pokusili zajistit co největší autentičnost textů, připouštějí ale, že v některých případech čerpali i z nespolehlivých moderních edic (Nevala & Nurmi, sek. 2.3, 2.4). Spolehlivost edic tedy bude při dalším výzkumu ještě třeba dále prověřit.

ANALÝZA

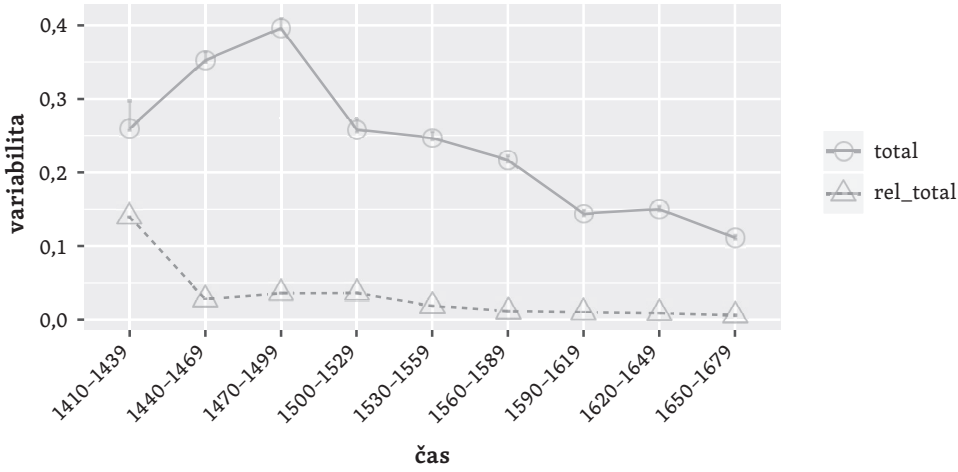
Prvním krokem analýzy bylo srovnání výše navržené míry variability založené na vážené entropii typů s mírou založenou na poměru forem a typů. Jak je patrné z grafu 4, obě míry ukazují klesající trend, respektive navzdory kolísání (v případě entropie) potvrzují vyšší variabilitu na počátku sledovaného období než na jeho konci. Jak již bylo řečeno, o postupném prosazování standardizace pravopisu v 15.–17. století nemůže být pochyb. Zajímavý je ale rozdíl vývoje především v prvních obdobích 1410–1469. Vyplývá z něj jednak to, že poměr forem a typů je výrazně závislý na velikosti vzorku — malé vzorky dat budou snadno tíhnout k vysoké variabilitě, což je další důvod, proč tuto míru nepoužívat.

V případě míry založené na vážené entropii je nízká variabilita, respektive její nárůst, vysvětlitelná obtížněji, svou roli bude ale zřejmě hrát kromě malého vzorku dat (viz značný rozsah konfidenčních intervalů a následující diskuse) i složení vzorku v tomto období (viz grafy 1–3): minimální zastoupení ženských autorů (4 %), převaha dopisů psaných písaři (75 %) a minimální zastoupení dopisů psaných v rámci rodiny (3 %). Jak ale vyplývá z další analýzy těchto proměnných, nelze těmito faktory vysvětlit překvapivě nízkou hodnotu v prvním období beze zbytku.

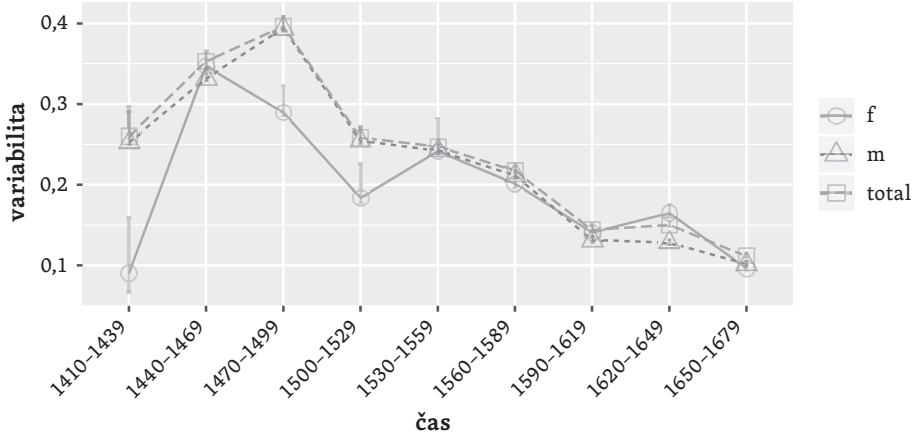
Další analýzou bylo srovnání míry variability podle pohlaví autorů zachycené grafem 5. Z něj vyplývá, že zhruba v první polovině celého sledovaného období (tedy do roku 1530) kolísá pravopis žen méně než pravopis mužů, ve druhé polovině je srovnatelný. V první polovině je ale zároveň zcela minimální zastoupení dopisů psaných přímo autorkami, neboť prakticky všechny dopisy žen jsou zapsány písaři (pravděpodobně tedy muži); viz k tomuto i graf 6 a následná diskuse.

Následně jsme srovnali variabilitu v závislosti na tom, kdo dopis skutečně zapsal (graf 6). Podle očekávání kolísal — kromě prvních dvou období — pravopis samotných autorů dopisů více než pravopis písařů.

6 Jedná se o odhady přesnosti výsledků, které jsou statisticky odvozené z dat pozorování. V grafech níže jsou značeny vertikálními úsečkami nad a pod datovými body.



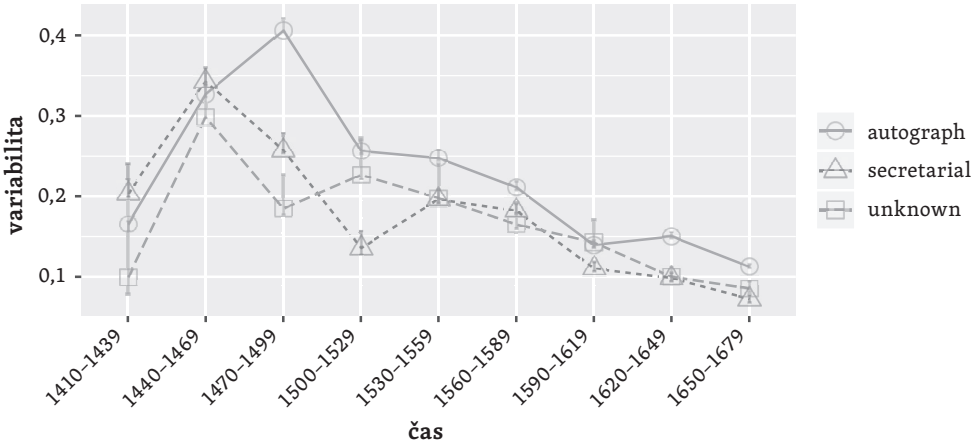
GRAF 4: Variabilita ortografie vyjádřená váženou entropií (total) a poměrem forem a typů (rel_total).



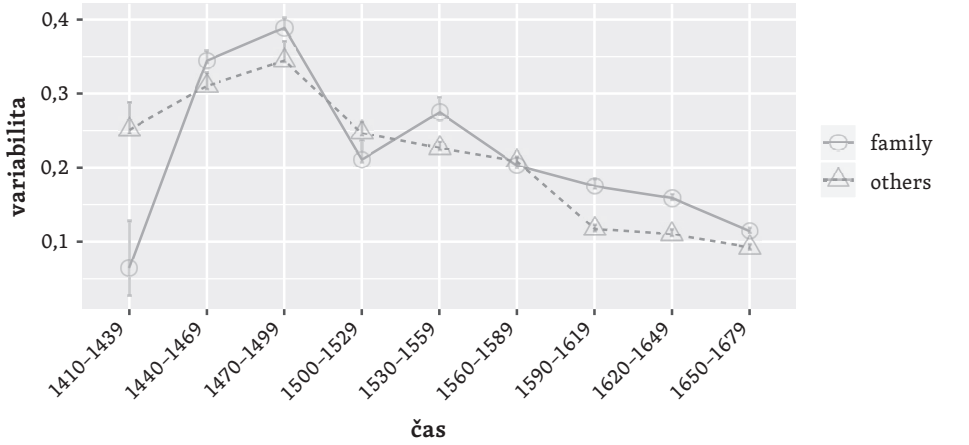
GRAF 5: Variabilita ortografie dle pohlaví autorů.

Jako třetí proměnnou jsme srovnali variabilitu v dopisech členům rodiny a ostatním (graf 7). Až na první období, kde šíře konfidenčních intervalů naznačuje minimální vzorek dat, a na jeden výkyv v letech 1500–1529 byl pravopis v dopisech psaných rodinným příslušníkům stejný nebo uvolněnější než v dopisech ostatním adresátům.

V posledních dvou částech analýzy jsme srovnávali souhru dvou proměnných. Graf 8 ilustruje zároveň vliv pohlaví a vliv autenticity na vývoj variability. Ukazuje, že rozdělení na malé kategorie přináší riziko malých vzorků dat. Nejenže zvlášť v raných obdobích jsou konfidenční intervaly velmi široké, ale v prvním období jsou také všechny kategorie autorek na nulové hodnotě, což se zdá odporovat grafu 5, kde je sice variabilita ortografie autorek v prvním období nízká, ale nenulová. Tímto pro-



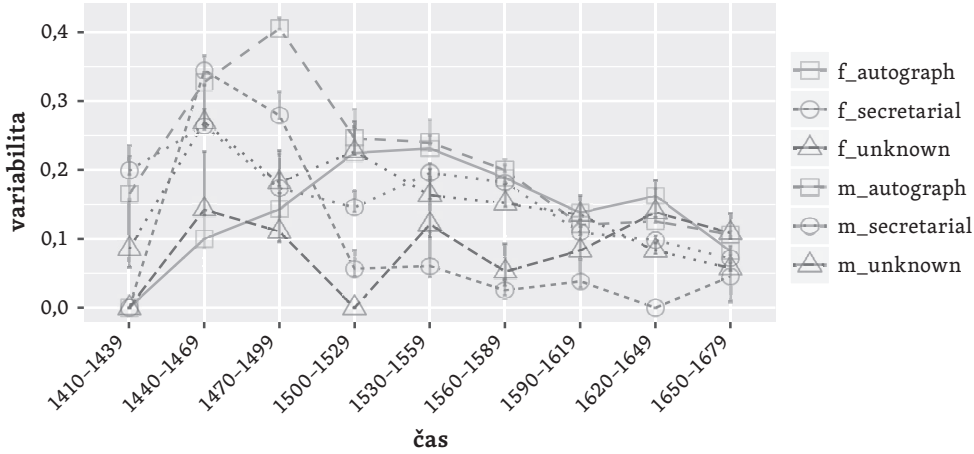
GRAF 6: Variabilita ortografie dle autenticity dopisu: dopisy psané přímo autory (autograph), písaři (secretarial) a nejistým pisatelem (unknown).



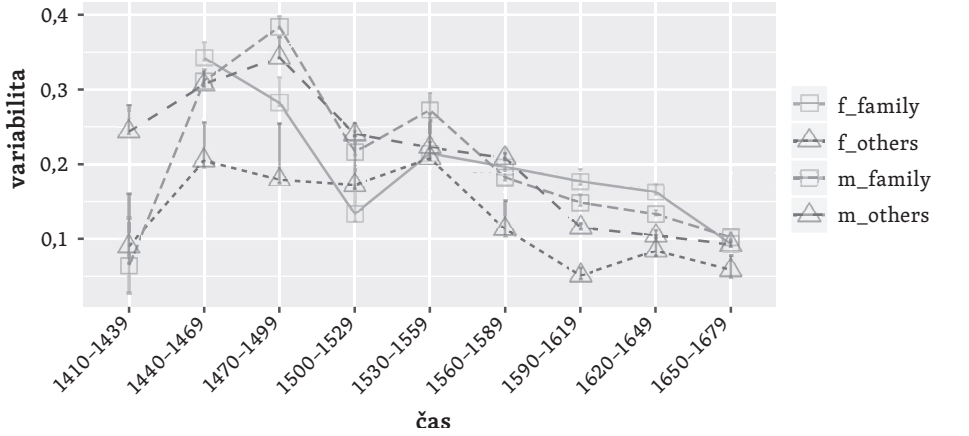
GRAF 7: Variabilita ortografie dle vztahu autora a adresáta: rodinný vztah (family) a ostatní (others).

blémem se budeme více zabývat v diskusi. Většina období ale napovídá, že rozdíl mezi variabilitou autorů a písařů (především u mužů) se postupně zužuje a rozdíl mezi autorkami a písaři je větší než v případě mužů.

Graf 9 srovnává zároveň vliv pohlaví a vliv vztahu k adresátovi. V případě mužů není ve variabilitě velký rozdíl, u žen vliv vztahu k adresátovi v první polovině sledovaného období také není jednoznačný, avšak ve druhé polovině je rozdíl, jak byl popsán pro graf 7, dobře patrný. Je také pozoruhodné, že v posledních čtyřech obdobích autorky píšící v rámci rodiny mají v pravopisu nejsilnější kolísání, naopak autorky píšící mimo rodinu mají ortografii nejstabilnější, tedy stabilnější než autoři. Toto pozorování dobře odpovídá tomu, co Labov nazývá „gender paradox", když říká, že „[w]omen conform more closely than men to sociolinguistic norms that are overtly prescribed, but conform less than men when they are not" (2001, s. 261–293).



GRAF 8: Ortografická variabilita dle kombinace proměnných pohlaví autora a autenticity dopisu.



GRAF 9: Ortografická variabilita dle kombinace proměnných pohlaví autora a vztahu k adresátovi.

DISKUZE

V průběhu analýzy jsme narazili na několik problémů, avšak nejvýraznější z nich bezesporu souvisí s velikostí vzorku. V diachronní korpusové lingvistice se samozřejmě jedná o problém všudypřítomný a entropie má alespoň teoretickou výhodu, protože v obecné rovině není závislá ani na velikosti, ani na typu vzorku. Výzkum variability ortografie však již z povahy věci nemůže být prováděn na příliš malých vzorcích — aby bylo možné kolísání pozorovat, je třeba, aby se slova, v našem případě typy, dostatečně často opakovala. Přestože námi analyzovaný vzorek nebyl úplně malý (téměř 300 tisíc tokenů), v případě rozdělení vzorku mezi různě zastoupená období a kategorie bylo nutné pracovat i se vzorky čítajícími několik málo desítek tokenů.

S tím se pojí další, dosud nezmíněná obtíž. Zkoumaný vzorek se neskládá z kompletních textů původních dopisů ani náhodně vybraných tokenů, ale z výskytů celkem jen 9 lexikálních jednotek, respektive 79 typů. Je nasnadě, že v malém vzorku, např. v textech autorek v prvním zkoumaném období, se několik málo výskytů zrovna těchto 9 lexikálních jednotek může psát shodně.

Zde se také musíme vrátit k problému naznačenému v analýze grafu 8: jak je možné, že v prvním zkoumaném období je entropie všech jednotlivých kategorií pro autorky nula, ale entropie pro autorky jako celek je dle grafu 5 nenulová. Je to dáno tím, že pokud se variantní výsledky rozdělí do více kategorií, v rámci těchto jednotlivých kategorií již nemusí k variabilitě docházet. Např. přítomné tvary třetí osoby slovesa BE jsou pro autorky prvního období celkem dvě, <is> a <ys>, variabilita je tedy nenulová. Každá forma se však vyskytuje v dopise rozdílné autenticity a rozdílného zařazení co do vztahu k adresátovi. V rámci těchto kategorií tedy bude daná variabilita nulová.

Jak velký by měl být nejmenší smysluplný vzorek pro práci s námi zavedenou mírou variability, není snadné jasně definovat, jelikož nejde jen o velikost vzorku, ale také o jeho lexikálně-gramatickou (tedy slovotvorně a gramaticky/flektivně morfologickou) různorodost. Obecně se však dle pozorovaných fluktuací ukazuje, že ve druhé polovině námi sledovaného období (ca od roku 1550) jsou již i vzorky za jednotlivé kategorie v třicetiletých úsecích dostatečně velké, pohybují se v řádech tisíců, zatímco vzorky ze začátku sledovaného období se pohybují jen v řádech desítek a jsou evidentně příliš malé.

S problematickou vypovídací hodnotou výsledků u malých vzorků dat jsme se vyrovnali výše popsanou metodou bootstrappingu — v grafech znázorněnou konfidenčními intervaly. Tím ale neřešíme problém rozdílné velikosti jednotlivých vzorků a jejich srovnání. Pokud narážíme na to, že v malém vzorku se vzhledem k nízkému zastoupení pozorovaných typů může entropie jevit nepřiměřeně nízká, naznačujeme, že entropie v jazyce obecně a zde specificky v ortografii zřejmě neroste lineárně s velikostí textu. Jde o poměrně intuitivní zjištění — v krátkém textu se variabilita nemá šanci projevit, naopak od určité délky textu zřejmě nebude příliš narůstat, protože variant zápisu nemůže být v přirozeném jazyce neomezeně mnoho.

Vidíme dvě možnosti, jak tento problém řešit. Buď z daného korpusu odvodit obecný vývoj růstu entropie vzhledem k velikosti vzorku a touto funkcí poté normalizovat výsledky našeho zkoumání, nebo korpus analyzovat výhradně po stejně velkých vzorcích. Druhá možnost již byla v korpusovém výzkumu implementována pomocí tzv. pohyblivého okna/úseku (např. Kubát & Milička, s. 341), kdy namísto po diskrétních úsecích daných stejně dlouhými časovými intervaly je korpus rozdělen na překrývající se okna/úseky definované shodným rozsahem tokenů.[7] Tento postup plánujeme v dalším výzkumu využít a doufáme, že jednak zpřesní výsledky pro méně zastoupené proměnné, jednak zvýší věrohodnost výsledků pro méně zastoupená období.

7 V případě srovnání souborů dat rozdělených na stejně velké celky, ale o rozdílné celkověvelikosti je samozřejmě třeba očekávat rozdílnou hustotu výsledných hodnot / datových bodů.

Dalším zajímavým problémem je kategorizace pohlaví autorů ve spojení s kategorií autenticity. Ač identitu, a tedy ani pohlaví písařů (nemusí jít nutně o profesionální písaře) neznáme, je velmi pravděpodobné, že řada, ne-li většina z nich byli muži. Pro určité druhy lingvistického výzkumu může být pohlaví autora důležitější než pohlaví samotného zapisovatele, avšak při výzkumu ortografie tomu tak není. Pracovat tedy s texty označenými v korpusu za autorsky ženské, ale zároveň písařské jako s texty žen je problematické. Zároveň je třeba připustit, že písaři mohou v tomto smyslu být i opisovači, což by ortografii autorek vracelo částečně zpět do hry.

ZÁVĚR

Výzkum si položil za cíl vytvořit novou metodologii kvantitativního měření ortografické variability na korpusovém základě, což se podařilo s využitím míry založené na vážené entropii. Navržená míra a metodologie umožňuje nový typ korpusového výzkumu, byť bude třeba výsledky této první sondy rozšířit na větší a lexikálně různorodější vzorek dat a srovnat je nejen s obdobným měřením starších a novějších období, ale i s měřením v korpusech neprivátních textů. Tak bude možné ukázat, zda postupné a setrvalé ubývání ortografické variability ještě v druhé polovině 17. století je specifikem dopisového žánru, nebo zda nově zavedená metodika naruší tradiční interpretaci o vývoji standardizace ortografie, podle níž se anglická ortografie ustálila víceméně již v textech 16. století (viz již výše citovaný Scragg, s. 46).

LITERATURA:

Canty, A., & Ripley, B. (2017). *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-20. Dostupné z: https://cran.r-project.org/web/packages/boot/

Crystal, D. (2005). *The Stories of English*. Woodstock, NY: The Overlook Press.

Cvrček, V. (2017). Variabilita. In P. Karlík, M. Nekula & J. Pleskalová (Eds.), *CzechEncy — Nový encyklopedický slovník češtiny*. Dostupné z: https://www.czechency.org/slovnik/VARIABILITA

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap Methods and Their Applications*. Cambridge: Cambridge University Press. Dostupné z https://cran.r-project.org/web/packages/boot/

Labov, W. (2001). *Principles of Linguistic Change, Vol. 2: Social Factors*. Malden, MA: Blackwell Publishers Inc.

Nevala, M., & Nurmi, A. (2013). The Corpora of Early English Correspondence (CEEC400). In A. Meurman-Solin & J. Tyrkkö (Eds.), *Principles and Practices for the Digital Editing and Annotation of Diachronic Data*. Tampere: University of Tampere. Dostupné z http://www.helsinki.fi/varieng/series/volumes/14/nevala_nurmi/

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*(3), 379–423.

Scragg, D. G. (1974). *A history of English spelling*. Manchester: Manchester University Press.

Upward, C., & Davidson, G. (2011). *The history of English spelling*. Malden, MA: Wiley-Blackwell.

KORPUSY:

Corpus of Early English Correspondence (CEEC) (1998). Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Jukka Keränen, Minna Nevala, Arja Nurmi and Minna Palander-Collin. Helsinki: University of Helsinki.

Parsed Corpus of Early English Correspondence, tagged version (PCEEC). (2006). Annotated by Arja Nurmi, Ann Taylor, Anthony Warner, Susan Pintzuk, and Terttu Nevalainen. Compiled by the CEEC Project Team. York: University of York and Helsinki.

The Helsinki Corpus of English Texts. (1991). Compiled by Matti Rissanen (Project leader), Merja Kytö (Project secretary); Leena Kahlas-Tarkka, Matti Kilpiö (Old English); Saara Nevanlinna, Irma Taavitsainen (Middle English); Terttu Nevalainen, Helena Raumolin-Brunberg (Early Modern English). Department of Modern Languages, University of Helsinki.

Ondřej Tichý | Ústav anglického jazyka a didaktiky FF UK
<ondrej.tichy@ff.cuni.cz>

PŘÍLOHA 1: SEZNAM TYPŮ, POČTŮ JEJICH FOREM A TOKENŮ

typ	počet forem	počet tokenů
BAG--	12	3356
BAG--1s	1	1
BE--	11	19067
BED--	3	5
BED--2s	2	5
BED--3s	2	2
BED--p	35	4579
BED--s	16	9714
BEI--	3	347
BEI--2s	1	1
BEN--	25	3992
BEP--	16	5658
BEP--1s	11	4981
BEP--2s	5	34
BEP--3s	15	19167
BEP--p	10	4771
BEP--s	5	8
DAG--	11	186
DAN--	20	1253
DO--	14	3698
DOD--	15	2310
DOD--2s	3	7
DOD--3s	1	2
DOI--	5	179
DOI--3s	2	2
DON--	17	1003
DOP--	10	3257
DOP--2s	4	17
DOP--3s	23	1200
EX--	20	3808
FOR--	5	1126
HAG--	28	1210
HAN--	7	341
HV--	10	217
HV--2s	1	1
HVD--	13	6090
HVD--2s	1	2
HVD--3s	1	3
HVD--p	1	1
HVI--p	3	71
HVN--	9	832
HVN--p	2	3
HV--p	18	7856
HVP--	13	67
HVP--2s	3	29
HVP--3s	25	7022
HVP--p	22	13068
MD-CAN-	16	5904
MD-CAN-2pret	1	2
MD-CAN-2s	1	3
MD-CAN-pret	38	2342
MD-DARE-	6	334
MD-DARE-3s	3	13
MD-DARE-pret	9	118
MD-MAY-	20	7362
MD-MAY-2pret	3	3
MD-MAY-2s	3	16
MD-MAY-pret	35	2331
MD-MUST-	18	2351
MD-NEED-	8	124
MD-NEED-3s	6	10
MD-NEED-pret	2	4
MD-OWE-	2	5
MD-OWE-3s	1	1
MD-OWE-pret	17	314
MD-SHALL-	39	10230
MD-SHALL-2pret	3	17
MD-SHALL-2s	3	22
MD-SHALL-p	3	4
MD-SHALL-pret	61	5582
MD-WILL-	40	11069
MD-WILL-2pret	1	2
MD-WILL-pret	54	5998
OTHER--	65	5001
SUCH--	49	4641
TO--	6	43000
TO-on-	2	9
WQ--	50	1085
WQ-if-	11	213