Radek Čech
Ioan-Iovitz Popescu
Gabriel Altmann

Edice Qfwfq

Metody kvantitativní analýzy (nejen) básnických textů

Olomouc
2014

Metody kvantitativní analýzy
(nejen) básnických textů
Radek Čech
Ioan-Iovitz Popescu
Gabriel Altmann

Recenzovali
PhDr. Ludmila Uhlířová, CSc., dr. h. c.
doc. Mgr. Ján Mačutek, Ph.D.

Tato publikace vychází v rámci grantu Inovace studia obecné jazykovědy a teorie komunikace ve spolupráci s přírodními vědami. reg. č. CZ.1.07/2.2.00/28.0076.

Tento projekt je spolufinancován Evropským sociálním fondem a státním rozpočtem České republiky.

1. vydání

ISBN 978-80-244-4044-6

OBSAH

1 Úvod

V průběhu posledních třiceti let došlo k poměrně významnému vývoji v oblasti kvantitativněligvistického bádání. Tento vývoj by se snad dal nejlépe charakterizovat jako hledání cest a způsobů, jak překonat popisný a klasifikační charakter analýz přirozeného jazyka. Jinými slovy, hlavním rysem tohoto vývoje je snaha formulovat zákony, jimiž se řídí jazykové chování, a hledat adekvátní metody, jimiž se platnost daných zákonů (a hypotéz s nimi souvisejících) empiricky testuje (více viz kap. 2).

Cílem této knihy je prezentovat některé z výsledků tohoto úsilí na příkladech analýzy básnických textů. V tomto smyslu je možné publikaci chápat jako další krok v tradici kvantitativních analýz poezie (srov. Altmann 1966; Altmann – Altmann 2008; Blatná 2001; Červenka – Sgallová 1997; Jakobson 1995; Levý 1964; Štukovský – Altmann 1964, 1965; Těšitelová 1968; Wimmer – Altmann – Hřebíček – Ondrejovič – Wimmerová 2003); jako krok, který by měl být vnímán především jako nastínění určitého způsobu bádání. K metodám zde představeným je nutné přistupovat kriticky, jsme přesvědčeni, že všechny mohou být modifikovány, vylepšeny, možná některé i odmítnuty jako neužitečné. Také se, jak doufáme, mohou stát inspirací pro rozvoj metod nových. To vše je ale možné jen za předpokladu, že budou široce aplikovány. Praxe však ukazuje, že největší potíž při aplikaci těchto a jim podobných kvantitativnělingvistických metod spočívá hlavně v určitém „strachu" z modelování jazyka prostřednictvím matematických a statistických nástrojů, který panuje mezi lingvisty a studenty lingvistických oborů, přičemž tento „strach" je v naprosté většině případů důsledkem neznalosti či předsudků. Svou roli samozřejmě hraje i neochota překonat uzavřený metodologický rámec oboru. Vzhledem k této skutečnosti jsme se pokusili napsat knihu, která představí aktuální stav určité části textové kvantitativní lingvistiky v co možná nejpřijatelnější podobě, tudíž – až na malé výjimky – nepředpokládá u čtenáře žádné předchozí znalosti matematiky ani statistiky (kromě některých poznatků nabytých na střední škole). A i v případě oněch výše zmíněných výjimek nic nebrání čtenáři použít navržený postup jako určitou

„kuchařku“, ve které sice úplně nebude rozumět jednotlivým krokům, ale bude stále vědět, jaký je smysl dané analýzy a jak ji interpretovat. Naší snahou bylo tedy prezentovat všechny metody co nejvíce „polopaticky“ – nejdříve je popsán každý krok analýzy a následně je vše ilustrováno na konkrétním příkladu.

Kniha je výsledkem naší několikaleté spolupráce, která se realizuje nikoliv jen při analýzách básnických textů. Proto je možné jednotlivé metody nalézt v dílčích studiích a článcích (Popescu – Čech – Altmann 2010, 2011a, 2011b, 2012a, 2012b, 2013; Čech – Popescu – Altmann 2011a, 2011b; Popescu – Altmann 2011), které však byly vzhledem k charateru knihy upraveny a rozšířeny. Kniha vznikla za podpory grantu „Lingvistická a lexikostatistická analýza ve spolupráci lingvistiky, matematiky, biologie a psychologie“, reg. č. CZ.1.07/2.3.00/20.0161.

Radek Čech, Ioan-Iovitz Popescu, Gabriel Altmann

2 Kvantitativní lingvistika

V českém lingvistickém prostředí má kvantitativní lingvistika (dále KL) poměrně dlouhou a bohatou tradici (srov. Uhlířová 2005). V rámci této tradice je kvantifikace jazykových jevů v naprosté většině případů používána jako doplňující charakteristika při popisu a klasifikaci jednotek jazykového systému (např. Trnka 1935, 1951; Krámský 1942; Jelínek – Bečka – Těšitelová 1961; Doležel 1963; Doležel – Průcha 1966; Těšitelová 1985, 1992; Čermák – Křen 2004; Čermák 2007; Bartoň – Cvrček – Čermák – Jelínek – Petkevič 2009).

Od 70. let 20 stol. se však pod stejným označením začíná formovat přístup, který kvantifikaci (a matematickou formalizaci) chápe jako nástroj *pro testování hypotéz* týkajících se vlastností jazykového systému (srov. Köhler 1986; Köhler – Altmann 2005, 2011; Köhler 2005a). Jinými slovy, KL tohoto typu si již neklade za cíl klasifikaci jazykových jevů (která je v lingvistice obvykle – podle nás omylem – považována za prostředek explanace), ale jejím cílem je modelovat vlastnosti jazykového systému a procesy s ním související tak, aby tyto vlastnosti bylo možné vyjádřit formou empiricky testovatelných hypotéz. Důležité při tom je, že hypotézy jsou zpravidla odvozovány buď z obecných principů, např. Zipfova principu nejmenšího úsilí – least effort principle (Zipf 1949), nebo (v lepším případě) z obecné jazykové teorie, např. synergetického jazykového modelu (Köhler 1986, 2005b). KL v tomto smyslu je tedy experimentální vědou stejně jako třeba chemie, fyzika či experimentální psychologie. A je to právě důsledné „lpění" na experimentálním přístupu (za použití běžně používaných statistických metod), jímž se KL tohoto druhu[1] liší od většiny ostatních lingvistických směrů.

Graficky bychom mohli (ideální) postup v KL znázornit takto:

1 V další části textu budeme termín „kvantitativní lingvistika" (KL) používat výhradně v tomto smyslu.

(teorie)
↓
verbální formulace hypotézy
↓
matematická formalizace
↓
experiment (konstrukce dat)
↓
matematické vyhodnocení experimentu (testování)
↓
lingvistická explanace

Vzhledem k tomu, že ne všechny hypotézy jsou formulovány ze vztahu k jazykové teorii, ale mohou vzniknout také jiným způsobem (srov. Bunge 1983; Feyerabend 2001), první krok jsme uzávorkovali. Připomeňme, že hypotézy odvozené z obecných principů nejrůznějšího druhu se mnohdy stávají inspirací při budování teorie.

2.1 SPECIFIKA KVANTITATIVNÍ LINGVISTIKY

Obecně lze říct, že cíle KL se na první pohled nijak výrazně neliší od cílů strukturální či generativní lingvistiky – smyslem bádání je co nejlépe poznat fungování jazyka. Na rozdíl od strukturalismu či generativismu však není cílem KL poznat pravidla jazyka, ale jazykové zákony. Dále, přijetí experimentální metodologie s sebou nevyhnutelně nese určitou změnu pohledu na jazyk a možnosti jeho zkoumání. Je-li totiž „úhelným kamenem" KL intersubjektivní empirický test (tj. experiment), není již možné se spoléhat na introspekci/intuici, která je ze své podstaty vždy subjektivní (srov. Estes 2000: 21): „With the decline of structuralism, introspective methods lost favor as a source of psychological data. Then, gone but not forgotten, introspection re-emerged in the 1960s with the rise of cognitive science, in which verbal protocols were a major source of data and a basis for much theorizing on problem solving. But despite their new popularity, introspective methods continued to exhibit the same weakness that had aroused critics in the structuralist period – *the lack of effective means of obtaining interpersonal agreement among scientists on the interpretation of introspective data*." (zvýraznili

Čech – Popescu – Altmann); ve stejném duchu o introspekci píše např. Meili (1967). Pokud ovšem odmítáme introspekci jako prostředek získávání a hodnocení jazykového materiálu (tradičně ve smyslu gramatický vs. negramatický, přijatelný vs. nepřijatelný), zůstává v současné době jediným přístupným zdrojem lingvistické analýzy jazykový projev (ať mluvený či psaný). To má ovšem vážný důsledek: jestliže je jediným zdrojem jazykový projev (de facto verbální chování), není cílem takového zkoumání poznat potencionalitu jazykového systému (tj. to, co je v jazyce možné – s podmínkou gramatické správnosti), nýbrž vytvářet modely, které se snaží postihnout a vysvětlit zákonitosti, jimiž se řídí jazykové chování. Přirozeným důsledkem tohoto postoje je odmítnutí dichotomického pohledu na jazyk – pojmy jako „langue", „competence" či „language faculty" nemají v tomto přístupu žádný praktický význam (srov. Čech 2005a). Ani v KL nikdo sice nepochybuje o lidské schopnosti (např. ve smyslu language faculty) používat jazyk, ale vzhledem k tomu, že není možné podrobit vlastnosti této schopnosti experimentálnímu zkoumání (máme na mysli tradiční lingvistické metody, nikoliv neurolingvistiku a neuropsychologii, která však spíše než vlastnosti nějakého jasně definovaného jazykového systému ve smyslu langue či competence zkoumá neurální koreláty jazyka), není předmětem zájmu KL. Navíc, dichotomický pohled na jazyk je z perspektivy vývoje evropského filozofického myšlení projevem metafyzického dualismu, typického např. pro platonismus, racionalismus či fenomenologii (srov. Čech 2005b, 2007).

Dalším typickým rysem KL je stochastický pohled na jazyk a jeho vlastnosti. Důvody jsou následující: vyjděme z předpokladu, že jazyk je systém, který je v nepřetržitém vývoji a podléhá neustálým změnám, které jsou důsledkem jazykové komunikace (srov. Bybee 2010; Hopper 1987; Hudson 2007; Kořenský 1987; Köhler 1986; Martinet 1964; Zipf 1935, 1949). Je přitom evidentní, že naprostá většina změn probíhá postupně: po první instanci následují další a další, pokud se změna „uchytí", nabírá na síle a v krajním případě se projeví jako jazykové pravidlo; z této perspektivy je tedy pravidlo v nejlepším případě jen extrémním případem stochastického zákona (tj. platí bez výjimky ve všech případech). Vzhledem k povaze jazykového vývoje se ale zdá rozumné předpokládat, že většina zákonů řídících jazykové chování nemá povahu pravidel, ale že

se projevují jako tendence. V důsledku toho jazykové zákony v KL nepostihují jednotlivé případy, ale predikují, s jakou pravděpodobností se určité jazykové jevy mohou vyskytnout za daných podmínek. Pro ilustraci tohoto přístupu vezměme dobře známý vztah mezi délkou slova a frekvencí, který lze formulovat takto: čím je slovo v textu frekventovanější, tím je kratší. Je evidentní, že téměř ve všech textech najdeme slova, která porušují tuto predikci (např. nalezneme čtyřslabičné slovo s frekvencí vyšší než slovo trojslabičné). Pokud ovšem bereme tuto predikci stochasticky (tj. jako tendenci), můžeme s explicitně definovanou pravděpodobností chyby (např. 5 %) testovat, zda je (či není) mezi délkou slova a frekvencí v daném textu vztah. Připomeňme, že stochastická povaha zákonů není jen doménou sociálních věd, kde se zdá být díky povaze předmětů zkoumání „přirozená", ale že stochastické modelování se ukázalo být velmi užitečným přístupem i v tzv. „tvrdých" vědách, jako je fyzika či chemie.

Se stochastickým pohledem na jazyk úzce souvisí i způsob klasifikace jazykových jednotek. Klasifikace je v rámci KL vnímána nikoliv jako cíl bádání, ale jako nezbytná podmínka pro možnost testování hypotéz. Je třeba zdůraznit, že v souladu s poznatky moderní filozofie vědy (Bunge 1983; Feyerabend 2001; Fraassen 2002; Polanyi 1962) se v rámci KL nepředpokládá, že by jazyková data byla nějak dopředu „dána" a že by těmto datům (s větší či menší přesností) odpovídaly naše pojmy. Naopak, jazykové jednotky jsou chápany jako naše konstrukce, prostřednictvím nichž se snažíme „manipulovat" s realitou (např. v rámci empirických experimentů). V tomto ohledu pak neexistuje nic jako lepší či horší klasifikace (ve smyslu pravdivější, tj. lépe odpovídající „objektivní" skutečnosti), ale jen klasifikace užitečnější či méně užitečná. A užitečnější je ta, která přináší lepší experimentální výsledky, samozřejmě takové, které se opírají o dobrá teoretická východiska a vedou k explanaci. Příkladem takové klasifikace jsou tzv. motivy (Köhler – Naumann 2008, 2010), což jsou jednotky v tradiční lingvistice zcela neznámé, a dá se říct, že do značné míry i neintuitivní. Smysluplnost používání motivů pak není v KL obhájena racionální diskusí, ale především prostřednictvím experimentů, které přinášejí stejně dobré (a někdy i lepší) výsledky než klasifikace založená na segmentaci jazykové promluvy například na slova.

Nyní zpět k samotným způsobům klasifikace.[2] Nejjednodušším způsobem třídění je metoda, která tvrzení o daném jazykovému jevu vzhledem ke zkoumané kategorii redukuje na dichotomické lišení pravda – nepravda. Máme-li tedy dva jazykové jevy *A* a *B*, můžeme vztah mezi nimi vzhledem ke zkoumané vlastnosti *V* vyjádřit formálně takto:

$V(A) = V(B)$, nebo $V(A) \neq V(B)$.

Připomeňme, že se jedná o způsob klasifikace, který je v lingvistice převládající (např. dělení na slovní druhy, větné členy, fráze, sémantické role atd.). Problém je v tom, že tento způsob zřejmě neodpovídá povaze jazykových jevů. Již Lakoff (1973) poukázal na to, že je adekvátnější jazykové jevy kategorizovat graduálně, tzn. některé jevy lze považovat za „lepší" reprezentanty dané kategorie než jiné (např. přímý akuzativ lze chápat jako typičtější případ objektu než předložkový pád). Následkem toho je pak možné jevy uspořádat ordinálně. Při tomto typu klasifikace se pak mohou mezi dvěma jazykovými jevy *A* a *B* vzhledem ke zkoumané vlastnosti *V* realizovat následující vztahy:

$V(A) > V(B)$, nebo $V(A) = V(B)$, nebo $V(A) < V(B)$.

Třetí způsob klasifikace je založen na kvantifikaci dané vlastnosti a následné možnosti jasného určení míry této vlastnosti vzhledem ke zkoumanému objektu. Máme-li tedy dva jazykové jevy *A* a *B*, můžeme vztah mezi nimi vzhledem k vlastnosti *V* formálně vyjádřit takto:

$V(A) - V(B) = d$,

kde *d* je číselná hodnota rozdílu.

Ilustrujme výhody třetího způsobu klasifikace na příkladu synonymie, která je původně čistě kvalitativním pojmem – slovo je tradičně označeno za synonymum, pokud je jeho význam stejný nebo podobný jako význam jiného slova. Pokud však synonymii kvantifikujeme, např. za použití slovníku synonym či tzv. synsetů (srov. Miller – Beckwith – Fellbaum – Gross – Miller 1993), můžeme

2 Podrobný popis způsobů klasifikace lze najít u Köhlera (2012, kap. 2.3.3), který byl hlavní inspirací této části; srov. také Čech (2013).

míru této vlastnosti vyjádřit pomocí celých čísel ležících v intervalu <0, ∞>. Je evidentní, že kvantifikace nabízí hlubší vhled do problematiky synonymie, a to i v případě setrvání u pouhého popisu tohoto jevu. Kvantifikace je v KL ale především nástrojem umožňujícím testovat hypotézy (viz výše). Konkrétně v případě synonymie je možné testovat vztah mezi takto kvantifikovanou synonymií a frekvencí, polysémií (také kvantifikovanou např. na základě počtu významů zaznamenaných ve výkladovém slovníku či synsetu), průměrnou délkou slov atd. (srov. Köhler 1986). Je snad na první pohled zřejmé, že tento způsob vede k takové interpretaci fungování synonymie, která není v původním čistě kvalitativním pojetí vůbec možná.

2.2 SHRNUTÍ

Základní charakteristiky přístupu KL a jejího pojetí jazyka lze shrnout do následujících bodů (srov. Altmann 2006; Altmann 2012).

(1) Jazyk je dynamický systém, jehož jednotky, vlastnosti, jednotlivé roviny, subsystémy a prostředí jsou ve vzájemném vztahu a ovlivňují se s různou mírou intenzity; pokud dojde ke změně např. jedné vlastnosti, projeví se to (s různou intenzitou) ve změnách jiných vlastností jazyka. Žádná jazyková vlastnost tedy není izolovaná.

(2) Jazyk má nekonečně mnoho vlastností. To, co nazýváme „vlastnostmi", není jazyku inherentní, ale jedná se o naše konstrukce.

(3) Vlastnosti jazyka jsou měřitelné; kvantifikace, která umožnuje měření, dovoluje postulovat přesnější definice a statisticky testovat hypotézy.

(4) Žádná jazyková vlastnost nemá nekonečný rozsah; tento princip je důležitý pro modelování (přestože můžeme nekonečný rozsah použít jako první aproximaci).

(5) Všechny vlastnosti jazyka podléhají změně. Toto je důsledkem používání jazyka, jiných změn jazykového systému (což se projevuje tzv. samoregulací) nebo požadavků (srov. termín „requirements" v synergetické lingvistice, viz Köhler (2005)) účastníků komunikace.

3 Tematická koncentrace textu

O naprosté většině textů (ať mluvených či psaných) můžeme říct, že se týkají nějakého tématu či souboru témat. Ze zkušenosti asi každý zná, že některé texty jsou tematicky vyhraněnější než jiné. Ilustrativní může být porovnání 1.–8. verše 3. kapitoly starozákonní knihy *Kazatel* s básní Bogdana Trojaka *Září pod Čantoryjí* (převzato z http://trojak.silesnet.com/index.php?text=ks28):

…
Všechno má svůj čas
Všechno má určenou chvíli
a veškeré dění pod nebem svůj čas:
Je čas rození i čas umírání,
čas sázet i čas trhat;
je čas zabíjet i čas léčit,
čas bořit i čas budovat;
je čas plakat i čas smát se,
čas truchlit i čas poskakovat;
je čas kameny rozhazovat i čas kameny sbírat,
čas objímat i čas objímání zanechat;
je čas hledat i čas ztrácet,
čas opatrovat i čas odhazovat;
je čas roztrhávat i čas sešívat,
čas mlčet i čas mluvit;
je čas milovat i čas nenávidět,
čas boje i čas pokoje.

Kazatel (3,1–8)

Září pod Čantoryjí
Beatě

Jako tehdy –
hřebenem farní střechy
vyčesat vesnici k Bohu,
co na horách roztáčí letokruhy;
ať je také dnes v hlavě
trocha zrychleného kolotání.

Dosud zde kolem chalup
běhá chlap se smrkovou latí,
odráží se, chce nahoru,
slzami roztlouká eternit nebes,
věčnosti obnažuje krovy.
Jako tehdy.

Jako tehdy čekat,
až půlnoc praskavě zakrouží zápěstím,
které ozdobil měsíc
věncem svých bílých kaplí,
a pak ti opět zašeptat:

a budeš mošt, jenž vytryskl
zpod křídel podzimního anděla…

Zatímco autor veršů knihy *Kazatel* se evidentně soustředí na jedno téma, tj. téma času a jeho atributů, báseň B. Trojaka je sledem lyrických obrazů a volných asociací, jimž dává určitou jednotu především název básně (je dost dobře možné si představit, že by se báseň jmenovala úplně jinak, čímž by se ale zřejmě radikálně změnilo i celé – tj. také tematické – vyznění). Na základě četby těchto dvou textů snad můžeme konstatovat, že *Kazatel* (3,1–8) je tematicky koncentrovanější než Trojakova báseň. Podobně jsou asi tematicky koncentrovanější např. mnohé odborné texty či novinové zprávy než umělecké eseje či povídky apod.

Abychom mohli překročit rámec intuitivního a subjektivního hodnocení této vlastnosti textu, tj. tematické koncentrace, je nezbytné tematické charakteristiky textu nějakým způsobem formalizovat. Jednou z možností je postup navržený Popescem (2007), rozpracovaný Popescem a kol.[i] (2009), Popescem – Altmannem (2011), Čechem – Popescem – Altmannem (2013) a dále aplikovaný Wilsonem (2009), Davidovou Glogarovou – Davidem – Čechem (2013), Čechem (2014a) a Davidovou Glogarovou – Čechem (2013); jedná se o přístup, který vychází z frekvenčních charakteristik jednotek textu a vlastnosti tzv. *h*-bodu (viz následující kapitola). Samozřejmě že existuje celá řada dalších způsobů, jak analyzovat tematickou charakteristiku textu, srov. zejména tzv. obsahovou analýzu (Neuendorf 2001; Krippendorff 2012).

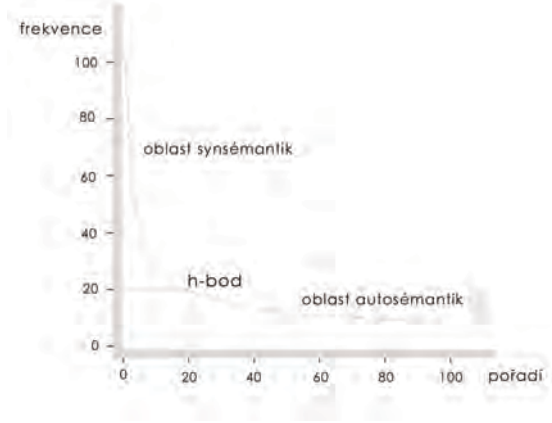
V následujících řádcích nejdříve představíme samotnou metodu měření tematické koncentrace (dále *TK*) (kap. 3.1), dále analyzujeme vztah mezi volbou jazykových jednotek a měřením *TK* (kap. 3.2), ukážeme, jak je možné rozdíly *TK* jednotlivých textů (případně skupin textů) statisticky testovat (kap. 3.3), a na závěr budeme krátce diskutovat vztah *TK* a dalších vlastností textu.

3.1 METODA MĚŘENÍ TEMATICKÉ KONCENTRACE TEXTU

Metoda měření *TK* textu je založena na vlastnostech frekvenční struktury textu, konkrétně na tzv. *h*-bodu (Popescu 2007, Popescu a kol.[i] 2009) a na pořadí a frekvenci slov vyjadřujících téma textu nad tímto bodem.[3]

3 Zavedení *h*-bodu v lingvistice bylo inspirováno Hirschovým indexem (*h*-indexem) používaným v scientometrii (Hirsch 2005), který je jedním z nejpoužívanějších indexů sloužících

Nejdříve obraťme pozornost k *h*-bodu, který hraje v této metodě rozhodující roli: seřadíme-li slova podle jejich klesající frekvence, tak *h*-bod je definován jako místo, v němž se pořadí slova rovná jeho frekvenci, srov. Obr. 3.1.



OBRÁZEK 3.1

h-bod oddělující dvě oblasti frekvenční distribuce slov; v grafu je hodnota *h*-bodu rovna 20, což znamená, že dvacáté nejfrekventovanější slovo v textu má frekvenci $f = 20$.

h-bod je tedy definován jako

$$h = \begin{cases} r, & \text{pokud } r = f(r) \\ \dfrac{f(i)r_j - f(j)r_i}{r_j - r_i + f(i) - f(j)} & \text{pokud } r \neq f(r) \end{cases}, \tag{3.1}$$

kde r je pořadí slova, $f(r)$ frekvence slova v daném pořadí, r_i a r_j jsou pořadí slov a $f(i)$ a $f(j)$ jsou jejich frekvence, přičemž $r_i < r_j$, kde r_i je největší takové číslo, pro které $r_i < f(i)$, a r_j je nejmenší takové číslo, pro které $r_j > f(j)$.

h-bod lze vnímat jako hranici, která ve frekvenční distribuci slov odděluje slova synsémantická (ta se obvykle vyskytují nad *h*-bodem, tj. jejich frekvence je vyšší než hodnota *h*-bodu: $f(r) > h$) od autosémantických. Samozřejmě že jde o hranici pouze přibližnou; v oblasti autosémantik se mohou vyskytnout

pro stanovení citačního ohlasu vědeckých článků, pomocí něhož lze hodnotit publikační aktivitu vědce.

synsémantika a naopak. Přítomnost autosémantických slov v oblasti synsémantik je tak možné považovat za jistý druh anomálie, která je odrazem specifické vlastnosti zkoumaného textu, konkrétně silné „zaměřenosti" (či „koncentrovanosti") autora na určité téma (či témata), reprezentované právě autosémantiky (či jinak vymezenými tzv. tematickými slovy, viz níže) nacházejícími se v oblasti synsémantik (srov. Obr. 3.1).

Vzhledem k tomu, že *h*-bod reprezentuje tzv. pevný bod ve frekvenční charakteristice textu, je možné jej použít jako východisko pro kvantifikaci tematické váhy slova (ať už slovo definujeme jako slovní formu, lemma, hreb či tzv. koreferenční jednotku, viz níže) a následně i pro kvantifikaci *TK* celého textu. Při stanovení indexu *TK* textu, prezentované Popescem a kol.[i] 2009, se bere v potaz jednak pořadí slov ve frekvenční distribuci, jednak jejich frekvence: vyjdeme-li z výše uvedeného předpokladu, že *TK* textu je reprezentována tematickými slovy nad *h*-bodem (označme pořadí těchto slov symbolem r'), je možné tematickou váhu těchto slov charakterizovat jako vzdálenost mezi *h*-bodem a pořadím slova nad *h*-bodem vynásobenou frekvencí tohoto slova, tj.

$$(h - r') \cdot f(r') \ . \tag{3.2}$$

Čím je tedy pořadí slova nižší (a čím je vyšší jeho frekvence), tím je jeho tematická váha větší. Abychom mohli tematické váhy použít pro porovnávání textů různé délky, je třeba je normalizovat. To uděláme tak, že každou hodnotu vypočítanou na základě vzorce (3.2) vydělíme sumou rozdílu vzdáleností $(h - r)$ u všech slov nad *h*-bodem a nejvyšší frekvencí slova v textu $f(1)$[4]; sumu všech rozdílů vzdáleností vypočítáme

$$\sum_{r=1}^{h} (h - r) = h\,(h) - \sum_{r=1}^{h} r = h^2 - \frac{h\,(h+1)}{2} = \frac{h\,(h-1)}{2} \ . \tag{3.3}$$

Vydělíme-li hodnotu vypočítanou na základě vzorce (3.2) touto sumou (3.3), můžeme stanovit index tematické váhy slova *TV* jako

4 Samozřejmě by bylo možné tematické váhy normalizovat i jinými způsoby, např. tak, že bychom použili sumu rozdílu vzdáleností $(h - r)$ u všech slov nad *h*-bodem a jejich frekvence.

$$(3.4) \quad TV_{slovo} = 2\frac{(h - r')f(r')}{h(h-1)f(1)} \ .$$

Tematická koncentrace celého textu je pak dána součtem hodnot tematických vah jednotlivých tematických slov *TV*, tj.

$$(3.5) \quad TK = \sum TV_{slovo} = \sum_{r'=1}^{T} 2\frac{(h - r')f(r')}{h(h-1)f(1)} \ ,$$

kde *T* je počet tematických slov nad *h*-bodem.

Podrobný postup měření *TK* budeme ilustrovat na textu *Kazatele* (3,1–8). Nejdříve je nutné si stanovit jazykové jednotky, které budou pro analýzu *TK* použity – pro jednoduchost začněme se slovními formami (o způsobech analýzy jiných jednotek viz kap. 3.2). Poté je třeba definovat vlastnosti slov reprezentujících tematické charakteristiky textu, tzv. tematická slova. Zde lze samozřejmě postupovat mnoha způsoby: např. je možné za reprezentanty tematických charakteristik brát všechna autosémantika nebo jen substantiva nebo substantiva, adjektiva a verba, jak to činí Popescu a kol.[i] (2009) atd. Stručně řečeno, stejně jako v případě volby jazykových jednotek (srov. kap. 3.2) neexistuje ani zde žádná a priori „správná" volba, vždy je třeba vycházet z potřeb dané analýzy. V našem případě budeme za tematická slova ve shodě s Popescem a kol.[i] (2009) považovat substantiva, adjektiva a verba (adjektiva a verba jsou vzhledem k substantivům predikáty prvního řádu).

Poté, co z textu vytvoříme frekvenční distribuci slov[5] (srov. Tab. 3.1), podle vzorce (3.1) vypočítáme hodnotu *h*-bodu. Vzhledem k tomu, že $r \neq f(r)$, použijeme druhou část vzorce, tj.

$$h = \frac{7 \cdot 4 - 2 \cdot 3}{4 - 3 + 7 - 2} = 3{,}67 \ .$$

5 Např. lze použít zdarma dostupný software Antconc (http://www.antlab.sci.waseda.ac.jp/software.html).

TABULKA 3.1
Frekvenční distribuce slov u *Kazatele* (3,1–8).

r	Slovo	*f*	*r*	Slovo	*f*	*r*	Slovo	*f*
1	*čas*	30	16	*milovat*	1	31	*roztrhávat*	1
2	*i*	14	17	*mluvit*	1	32	*sbírat*	1
3	*je*	7	18	*mlčet*	1	33	*se*	1
4	*kameny*	2	19	*nebem*	1	34	*sešívat*	1
5	*má*	2	20	*nenávidět*	1	35	*smát*	1
6	*svůj*	2	21	*objímat*	1	36	*sázet*	1
7	*všechno*	2	22	*objímání*	1	37	*trhat*	1
8	*a*	1	23	*odhazovat*	1	38	*truchlit*	1
9	*boje*	1	24	*opatrovat*	1	39	*umírání*	1
10	*bořit*	1	25	*plakat*	1	40	*určenou*	1
11	*budovat*	1	26	*pod*	1	41	*veškeré*	1
12	*chvíli*	1	27	*pokoje*	1	42	*zabíjet*	1
13	*dění*	1	28	*poskakovat*	1	43	*zanechat*	1
14	*hledat*	1	29	*rození*	1	44	*ztrácet*	1
15	*léčit*	1	30	*rozhazovat*	1			

Jak je vidět v Tab. 3.1, jediným tematickým slovem nad *h*-bodem je slovo *čas*. Na základě vzorce (3.4) můžeme určit jeho tematickou váhu, tj.

$$TV_{čas} = 2\frac{(3{,}67-1)\,30}{3{,}67\,(3{,}67-1)\,30} = 0{,}54496\ .$$

Jelikož se nad *h*-bodem nenachází žádné jiné tematické slovo, je hodnota *TK* celého textu *Kazatele* (3,1–8) rovna hodnotě $TV_{čas}$, tj. $TK_{Kazatel} = 0{,}54496$, což představuje extrémně vysokou hodnotu (většinou se hodnota *TK* pohybuje v řádu setin či tisícin). Naproti tomu v básni *Září pod Čantoryjí* se nad *h*-bodem ($h = 2{,}5$) nenachází žádné tematické slovo (srov. Tab. 3.2), proto je hodnota její *TK* rovna nule.

TABULKA 3.2
Frekvenční distribuce slov v básni *Září pod Čantoryjí*.

r	Slovo	*f*	*r*	Slovo	*f*	*r*	Slovo	*f*
1	*jako*	3	24	*jenž*	1	47	*roztlouká*	1
2	*tehdy*	3	25	*k*	1	48	*roztáčí*	1

r	Slovo	f	r	Slovo	f	r	Slovo	f
3	*a*	2	26	*kaplí*	1	49	*slzami*	1
4	*se*	2	27	*kolem*	1	50	*smrkovou*	1
5	*anděla*	1	28	*kolotání*	1	51	*střechy*	1
6	*ať*	1	29	*krovy*	1	52	*svých*	1
7	*až*	1	30	*které*	1	53	*také*	1
8	*bohu*	1	31	*křídel*	1	54	*ti*	1
9	*budeš*	1	32	*latí*	1	55	*trocha*	1
10	*bílých*	1	33	*letokruhy*	1	56	*v*	1
11	*běhá*	1	34	*mošt*	1	57	*vesnici*	1
12	*chalup*	1	35	*měsíc*	1	58	*vytryskl*	1
13	*chce*	1	36	*na*	1	59	*vyčesat*	1
14	*chlap*	1	37	*nahoru*	1	60	*věncem*	1
15	*co*	1	38	*nebes*	1	61	*věčnosti*	1
16	*dnes*	1	39	*obnažuje*	1	62	*zakrouží*	1
17	*dosud*	1	40	*odráží*	1	63	*zašeptat*	1
18	*eternit*	1	41	*opět*	1	64	*zde*	1
19	*farní*	1	42	*ozdobil*	1	65	*zpod*	1
20	*hlavě*	1	43	*pak*	1	66	*zrychleného*	1
21	*horách*	1	44	*podzimního*	1	67	*zápěstím*	1
22	*hřebenem*	1	45	*praskavě*	1	68	*čekat*	1
23	*je*	1	46	*půlnoc*	1			

3.2 ZPŮSOBY MĚŘENÍ TEMATICKÉ KONCENTRACE TEXTU – VOLBA JAZYKOVÝCH JEDNOTEK

Pravděpodobně jednou z prvních námitek proti metodě popsané v předchozí kapitole by byla námitka týkající se volby jazykových jednotek. Zejména u jazyka s bohatou flexí, jako je čeština, není volba slovních tvarů jako nositelů tematických charakteristik textu na první pohled volbou nejvhodnější. Na její obhajobu bychom mohli uvést fakt, že distribuce slovních tvarů je v rámci jednotlivých lemmat pravidelná, tudíž pro porovnávání tematických koncentrací jednotlivých textů není ani volba slovních tvarů metodou a priori „špatnou". Ale obecně vzato

je u silně flektivních jazyků eliminace vlivu flexe žádoucí, tudíž se jako vhodnější jednotky jeví např. lemmata, tj. základní podoby lexému reprezentující všechny tvary daného lexému (pro substantiva, adjektiva, zájmena a číslovky jde o nominativ singuláru maskulina; pro verba infinitiv; pro adverbia pozitiv; přičemž pro všechny případy zpravidla platí, že jsou do nich zahrnuty i negované formy).

S problémem volby jazykových jednotek v případě analýzy *TK* souvisejí otázky týkající se určování jazykových jednotek obecně. V první řadě je třeba si uvědomit, že neexistuje nic jako „přirozená" jazyková jednotka (a to se týká všech jazykových rovin). To, co se zpravidla označuje jako fonémy, morfémy či morfy, slova, lemmata, klauze, věty atd., není nic jiného, než *námi vytvořené* nástroje, pomocí nichž se snažíme zachytit vlastnosti toho, co nazýváme jazykem. Znamená to tedy, že je volba jazykových jednotek zcela libovolná? Nikoliv. Bezpochyby existují jednotky, které se pro danou analýzu hodí lépe než jednotky jiné. Obecně bychom tedy mohli říct, že neexistují jednotky „lepší" ani „horší" (ve smyslu „pravdivější"), ale jednotky „užitečnější" a „méně užitečné". Míra užitečnosti pak vždy záleží na badatelském cíli (např. testování hypotéz nejrůznějšího druhu, určování autorství, žánrových charakteristik atd.).

Dále je třeba si uvědomovat omezení, která s sebou volba jednotlivých jednotek přináší. Např. použijeme-li pro analýzu *TK* lemmata, musíme znát povahu lemmatizace – srov. rozdílnost lemmatizace použité v Českém národním korpusu (SYN 2013), která je založena pouze na formálních vlastnostech slov, tj. nerespektuje polysémii, s lemmatizací Pražského závislostního korpusu 2.0 (Hajič a kol. 2006), která se polysémii snaží do jisté míry respektovat (vinou čehož ale na druhé straně roste její chybovost). Pokud bychom chtěli při analýze *TK* respektovat i koreferenční vztahy mezi jednotkami označujícími stejné entity (všeho druhu), což se jeví jako jeden z nejvhodnějších způsobů, narážíme zase na omezené možnosti automatické detekce těchto vztahů – jejich postižení prostředky automatické analýzy textu je jednak relativně velmi náročné, jednak dosud i dosti nedokonalé. Proto se jako nejvhodnější řešení v tomto případě nabízí ruční anotace koreference, ta je však na druhou stranu devalvována častou absencí tzv. mezianotátorské shody – různí lidé anotují texty různými

způsoby. Jak je vidět, každá metoda má svá pro a proti a o žádné z nich nelze tvrdit, že je nejlepší.

V následující části představíme, jak se hodnota *TK* mění právě v závislosti na volbě jazykových jednotek. Konkrétně budeme analyzovat *TK* básně E. Bachletové *Iba neha*, a to třemi způsoby: 1) prostřednictvím slovních tvarů, 2) lemmat a 3) tzv. hrebů.[6] Původní text je zde:

Iba neha

Počítam s tvojím hlasom
tvojou nehou
tvojím slovom
a som prekvapená
ako ľahko sa
stávam závislá
na niečom
tak neuveriteľne
neskutočnom
závratnom
na niečom
čo sa bojím
bližšie označiť
bližšie skúmať
lebo obaja
dávno vieme
že ide o nás
a o veľa.

….-….

Dotýkaš sa ma
slovami
hlasom
perami
a ja cítim
že vo mne prebúdzaš
ženu
lásku
nádej
čakanie
a je mi
tak zvláštne
a neuveriteľne
dobre.

….-….

Tíšiš ma
a ja už cítim
to objatie
v ktoré dúfam
v ktoré dúfame.

….-….

A som s tebou spojená
spätá
uväznená v láske
som zovretá
v tichu, ktoré sa otvorí
keď už nemôžeme povedať
viac
než: ľúbim ťa.

….-….

A neviem čo príde
a neviem či prídeš
a neviem či tu ešte budem
či tu – ešte budeme
a v tom všetkom
zneistení
plačem, smejem sa
a dúfam, že naša láska
toto všetko unesie.

6 Hreb je nadvětná jazyková jednotka navržená L. Hřebíčkem (srov. Hřebíček 1997, 2002), který ji původně označil termínem agregát. Někdy je hreb označován také termínem sémantický konstrukt (srov. Andres – Benešová 2011, Benešová 2011).

Tab. 3.3 obsahuje frekvenční distribuci deseti nejfrekventovanějších slovních tvarů v básni *Iba neha*. Vzhledem k tomu, že některá slova mají stejnou frekvenci, je pro správný výpočet *TK* nejdříve nutné určit průměrné pořadí těchto slov (srov. druhý sloupec Tab. 3.3). Např. slova *či, ktoré, neviem, som, že* mají stejnou frekvenci $(f = 3)$, proto je jejich pořadí v prvním sloupci Tab. 3.3 do jisté míry zavádějící. Průměrné pořadí vypočítáme podle vzorce (3.6), tj.

$$\overline{r(f_i)} = \frac{\sum r(f_i)}{N(f_i)} \quad , \tag{3.6}$$

kde $r(f_i)$ je pořadí slov se stejnou frekvencí a $N(f_i)$ je počet těchto slov. V případě slov s $f = 3$ z Tab. 3.3 dostáváme

$$\overline{r(f_i)} = \frac{4+5+6+7+8}{5} = 6 \ .$$

TABULKA 3.3
Frekvenční distribuce deseti nejfrekventovanějších slovních tvarů v básni *Iba neha*.

r	Průměr (*r*)	Slovní tvar	*f*
1	1	*a*	12
2	2,5	*sa*	5
3	2,5	*v*	5
4	6	*či*	3
5	6	*ktoré*	3
6	6	*neviem*	3
7	6	*som*	3
8	6	*že*	3
9	17	*bližšie*	2
10	17	*cítim*	2

Hodnota *h*-bodu (podle vzorce (3.1)) je

$$h_{slovní\ formy} = \frac{5 \cdot 4 - 3 \cdot 3}{4 - 3 + 5 - 3} = 3{,}67 \quad .$$

Jelikož se nad *h*-bodem nevyskytuje žádné tematické slovo, je hodnota *TK* celé básně rovna nule.

Pokud text lemmatizujeme, získáváme údaje uvedené v Tab. 3.4.

TABULKA 3.4
Frekvenční distribuce deseti nejfrekventovanějších lemmat v básni *Iba neha*.

r	Průměr (*r*)	Lemma	*f*
1	1	*a*	12
2	2,5	*byť*	6
3	2,5	*v*	6
4	4,5	*ja*	5
5	4,5	*ty*	5
6	6	*vedieť*	4
7	9	*či*	3
8	9	*dúfať*	3
9	9	*ktorý*	3
10	9	*láska*	3

Hodnota *h*-bodu je $h = 5$ a stejně jako v předchozím případě se nad ním nenachází žádné tematické slovo, jak jsme si jej definovali výše. Na druhou stranu je celkem překvapivé, že se nad *h*-bodem nacházejí zájmena *ja* a *ty*. Vzhledem k tomu, že obě referují vždy k jednomu jasně vymezenému referentu (*ja* k autorce a *ty* k milenci), je možné je uznat za nositele tematické charakteristiky textu (netřeba snad zdůrazňovat, že v případě porovnávání různých textů je třeba být při definování tematických slov konzistentní). V tomto případě je hodnota *TK* celé básně

$$TK_{lemmata} = TV_{ja} + TV_{ty} = 2\frac{(5-4{,}5)\,5}{5\,(5-1)\,12} + 2\frac{(5-4{,}5)\,5}{5\,(5-1)\,12} = 0{,}020833\ .$$

Význam zájmen *ja* a *ty* však ve slovenštině nemusí být obligatorně vyjádřen pouze formou zájmena samotného, ale realizuje se i prostřednictvím koncovky, která mimochodem vyjadřuje nejen osobu, ale i jiné gramatické kategorie (proto je slovenština řazena mezi tzv. fúzní jazyky, srov. Comrie 1989). Pokud tedy chceme získat přesnější obraz o *TK* této básně (a pokud definujeme zájmena jako tematická slova), musíme započítat nejen zájmena samotná, ale také koncovky

sloves vyjadřující osobu, tj. *neviem, som, cítim, dúfam, bojím sa, budem, ľúbim, plačem, počítam, smejem sa, stávam sa.* Navíc je reference k autorce vyjádřena také v některých plurálových tvarech, proto je třeba započítat rovněž slovesa *budeme*, *dúfame*, *nemôžeme*, *vieme*, zájmena *nás*, *náš* a číslovku *obaja.* Tento způsob zpracování jazykových jednotek, tzv. hrebů, je detailně představen Zieglerem a Altmannem (2002) a bývá označován jako hrebová analýza. Stručně řečeno, hreb je jazyková jednotka odkazující ke stejné entitě. Proto se v hrebové analýze neberou v potaz předložky (sémanticky se dají chápat jako součást substantiva) či spojky (které mají pouze gramatický význam). Frekvenční distribuce deseti nejfrekventovanějších hrebů v básni *Iba neha* je uvedena v Tab. 3.5.

TABULKA 3.5
Frekvenční distribuce deseti nejfrekventovanějších hrebů v básni *Iba neha*. Pro snazší identifikaci je za jednotlivými elementy uvedeno pořadí daného elementu v básni (číslo označuje pozici slova).

r	Průměr (*r*)	Hreb	Elementy	*f*
1	1	*ja*	{počíta**m** 1, so**m** 10, stáva**m** sa 14–15, bojí**m** sa 26–27, obaja 33, vie**me** 35, nás 39, ma 45, ja 50, cíti**m** 51, mne 54, mi 62, ma 69, ja 71, cíti**m** 73, dúfa**m** 78, dúfa**me** 81, so**m** 83, so**m** 91, nemôže**me** 100, ľúbi**m** 104, nevie**m** 107, nevie**m** 111, nevie**m** 115, bude**m** 119, bude**me** 123, plače**m** 129, smeje**m** sa 130–131, dúfa**m** 133, naša 135}	30
2	2	*ty*	{tvojím 3, tvojou 5, tvojím 7, obaja 33, vie**me** 35, nás 39, prebúdza**š** 55, tíši**š** 68, dúfa**me** 81, tebou 85, nemôže**me** 100, ťa 109, príde**š** 113, bude**me** 123, naša 135}	15
3	3	*byť*	{som 10, je 61, som 83, som 91, budem 119, budeme 123}	6
4	4	*my*	{obaja 33, nás 39, dúfa**me** 81, nemôže**me** 100, naša 135}	5
5	5,5	*vedieť*	{vieme 35, neviem 107, neviem 111, neviem 115}	4

r	Průměr (*r*)	Hreb	Elementy	*f*
6	5,5	*všetko*	{tom 126, všetkom 127, toto 137, všetko 138}	4
7	7,5	*láska*	{lásku 57, láske 90, láska 136}	3
8	7,5	*objatie*	{to 74, objatie 75, ktoré 80, ktoré 95}	3
9	15,5	*slovo*	{slovom 8, slovami 46}	2
10	15,5	*hlas*	{hlasom 4, hlasom 47}	2

Hodnota *h*-bodu je $h = 4{,}5$. Jak je patrné z Tab. 3.5, tematickými hreby nad *h*-bodem jsou hreby {*ja*}, {*ty*} a {*my*}. Celková *TK* básně při hrebové analýze je

$$\begin{aligned} TK_{hreby} &= TV_{\{ja\}} + TV_{\{ty\}} + TV_{\{my\}} = \\ &= 2\frac{(4{,}5-1)\,30}{4{,}5\,(4{,}5-1)\,30} + 2\frac{(4{,}5-2)\,15}{4{,}5\,(4{,}5-1)\,30} + 2\frac{(4{,}5-4)\,5}{4{,}5\,(4{,}5-1)\,30} = \\ &= 0{,}613757\ , \end{aligned}$$

což je hodnota téměř třicetkrát větší než hodnota *TK* vypočítaná prostřednictvím analýzy lemmat.

3.3 TESTOVÁNÍ ROZDÍLŮ TEMATICKÉ KONCENTRACE

Rozdílné hodnoty *TK* jednotlivých textů, případně celých slupin textů mohou být projevem různých faktorů, jako jsou autorství, žánr, ideologická zatíženost atd. (srov. Čech 2014a; Davidová Glogarová – David – Čech 2013; Davidová Glogarová – Čech 2013). Pro adekvátní porovnání těchto hodnot je třeba rozdíly statisticky testovat. V následujících řádcích proto detailně popíšeme, jak postupovat při porovnávání *TK* jednotlivých textů, přičemž celý postup bude ilustrován na porovnání *TK* dvou básní J. Nerudy z *Písní kosmických*.

Pro aplikaci statistického testu je nutné znát nejen hodnoty *TK*, ale také rozptyl *TK* jednotlivých textů. Popescu a Altmann (2011) odvodili vzorec pro výpočet rozptylu *TK*

$$(3.7) \quad Var\,(TK) = \left(\frac{2}{h\,(h-1)\,f(1)}\right)^2 \cdot \left(\sum_{r'=1}^{T} f(r')\right) \cdot m_{2r'} \ ,$$

kde $m_{2r'}$ je rozptyl (druhý centrální moment) tematických slov nad h-bodem, tj.

$$(3.8) \quad m_{2r'} = \frac{\sum_{r'=1}^{T} (r' - m_{1r'})^2 f(r')}{\sum_{r'=1}^{T} f(r')} \ ,$$

kde $m_{1r'}$ je první počáteční moment, tj.

$$(3.9) \quad m_{1r'} = \frac{\sum r' \cdot f(r')}{\sum f(r')} \ .$$

Rozdíly hodnot *TK* jednotlivých textů porovnáme prostřednictvím asymptotického *u*-testu[7],

$$(3.10) \quad u = \frac{TK_1 - TK_2}{\sqrt{Var\,(TK_1) + Var(TK_2)}} \ .$$

Pokud chceme porovnávat skupiny textů, použijeme průměrné hodnoty *TK* a do jmenovatele vzorce (3.10) dosadíme namísto *Var* (*TK*) hodnotu podílu rozptylu průměrů tematické koncentrace s^2 a počtu měření *n*, tj.

$$(3.11) \quad u = \frac{\overline{TK}_1 - \overline{TK}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \ .$$

Celý postup nyní ilustrujme prostřednictvím porovnání lemmatizované osmé (*Poeto Světe…*) a třicáté páté (*Přijdou dnové…*) básně *Písní kosmických* J. Nerudy. Frekvenční distribuce lemmat obou básní jsou v Tab. 3.6 a 3.7.

7 Ve statistice je označován také jako jako *z*-test.

TABULKA 3.6
Frekvenční distribuce deseti nejfrekventovanějších lemmat v básni J. Nerudy *Poeto Světe…*

r	Průměr (*r*)	Lemma	*f*
1	1	*být*	31
2	2	*on*	17
3	3	*v*	16
4	4	*a*	14
5	5	***poeta***	9
6	6	*se*	8
7	7,5	*co*	7
8	7,5	*hymnus*	7
9	9,5	*svět*	6
10	9,5	*tvůj*	6

TABULKA 3.7
Frekvenční distribuce dvanácti nejfrekventovanějších lemmat v básni J. Nerudy *Přijdou dnové…*

r	Průměr (*r*)	Lemma	*f*
1	1	***země***	9
2	3,5	*být*	6
3	3,5	*dávno*	6
4	3,5	*v*	6
5	3,5	***věk***	6
6	6	*se*	5
7	7	*a*	4
8	10	*i*	3
9	10	*jak*	3
10	10	*k*	3
11	10	*letět*	3
12	10	*po*	3

Na základě postupu prezentovaného v předchozí kapitole vypočítáme hodnoty *TK* obou básní, tj.

$$TK_{Poeto\ světe\ldots} = 0{,}027650\ ,$$

$$TK_{Přijdou\ dnové\ldots} = 0{,}471381\ .$$

Jelikož se v básni *Poeto Světe…* vyskytuje pouze jedno tematické slovo, je rozptyl *TK* roven nule. V případě básně *Přijdou dnové…* provedeme výpočet rozptylu následujícím způsobem:

$$m_{1r'(Přijdou\ dnové\ldots)} = \frac{1 \cdot 9 + 3{,}5 \cdot 6}{9 + 6} = 2\ ,$$

$$m_{2r'(Přijdou\ dnové\ldots)} = \frac{(1-2)^2 \cdot 9 + (3{,}5-2)^2 \cdot 6}{9 + 6} = 1{,}5\ ,$$

$$Var\,(TK_{Přijdou\ dnové\ldots}) = \left(\frac{2}{5{,}5 \cdot 4{,}5 \cdot 9}\right)^2 \cdot\ (9+6) \cdot 1{,}5 = 0{,}001814\ .$$

Na základě vzorce (3.10) vypočítáme hodnotu testového kritéria *u*

$$u = \frac{|0{,}027650 - 0{,}471381|}{\sqrt{0 + 0{,}001814}} = 10{,}42\ ,$$

což je vysoce signifikantní rozdíl, protože vypočítaná hodnota *u* je výrazně vyšší než 1,96 (na hladině významnosti $\alpha = 0{,}05$ je rozdíl signifikantní, jestliže $u > 1{,}96$). Můžeme tedy tvrdit, že obě básně jsou vzhledem k hodnotě *TK* významně rozdílné.

3.4 TEMATICKÁ KONCENTRACE A JINÉ VLASTNOSTI TEXTU

Stejně jako ostatní vlastnosti jazyka a textu není ani *TK* vlastností izolovanou. Dá se předpokládat, že bude ve vztazích s jinými vlastnostmi reflektujícími tematické charakteristiky textu, zejména tzv. slovním bohatstvím (srov. kap. 4), a dále s pragmatickými faktory, jako jsou žánr, autorství, doba atd. Hlubší vhled do fungování *TK* mohou přinést analýzy zaměřené na testování následujících hypotéz:

- čím nižší slovní bohatství textu, tím vyšší *TK*; při formulování této hypotézy vycházíme z předpokladu, že čím více nových slov (lemmat či hrebů) autor použije, tím větší bude počet témat, kterými se bude v textu zabývat, což by se mělo odrazit na nižší hodnotě *TK* celého textu;
- čím vyšší index opakování slov (srov. kap. 4.2), tím vyšší *TK*; zdá se snad rozumné předpokládat, že větší opakování slov (lemmat či hrebů) by se mělo odrazit na vyšší hodnotě *TK* celého textu;
- čím nižší entropie (srov. kap. 4.3), tím vyšší *TK*; strukturovanější text – vzhledem k distribuci slov (lemmat či hrebů) – by se měl projevit ve vyšší hodnotě *TK* textu.

Je třeba zdůraznit, že některé z těchto hypotéz jistě budou muset být modifikovány, případně budou vyvráceny, protože indexy slovního bohatství, opakování slov a entropie charakterizují vlastnost celé frekvenční struktury textu, nikoliv jen části nad *h*-bodem, jak to činí *TK*. Nejdříve je proto třeba zjistit, zda tyto hypotézy vyjadřují nějakou obecnou tendenci, a následně podrobněji analyzovat příčiny porušování této tendence.

4 Slovní bohatství textu

Měření tzv. slovního bohatství má poměrně dlouhou tradici. Při analýze tohoto fenoménu je ovšem důležité si uvědomit především dvě věci: (1) jedná se o vlastnost, kterou jsme definovali konceptuálně my, abychom postihli něco, čím se jednotlivé texty odlišují, tj. není to něco, co by bylo v textu něčím reálným; (2) v důsledku toho je možné slovní bohatství kvantifikovat mnoha různými indikátory, přičemž není dáno žádné „přirozené" kritérium, jež by rozhodlo, který způsob měření je lepší než jiný. Je možné uvažovat pouze o tom, že „lepší" je takové měření, které nejlépe odpovídá nějakému zákonu. Pro vysvětlení fungování bohatství textu však prozatím žádný zákon nemáme – jsme ve stadiu hledání a objevování určitých pravidelností a je samozřejmě otevřenou otázkou, k jakým výsledkům v tomto směru bádání nakonec dospějeme.

Pokud má být měření slovního bohatství smysluplné, např. pro porovnání autorských či žánrových charakteristik, je třeba nějakým způsobem eliminovat vliv délky textu na hodnotu indexu vyjadřujícího slovní bohatství. To, že s rostoucí délkou textu roste i velikost slovníku, je evidentní. Je také evidentní, že jen málokdy máme příležitost porovnávat texty stejné délky. Asi nejjednodušším způsobem, jak celý problém vyřešit, je analyzovat pouze části textu, např. počátečních sto či tisíc slov – srov. nastavení WordSmith Tool pro analýzu poměru počtu výskytů různých slovních tvarů (types) a počtu všech slov v textu (tokens) (Scott 2011). Tento přístup je ovšem v mnohých ohledech neadekvátní. Zejména proto, že nebere v potaz tzv. homogenitu textu: představme si například detektivní povídku, ve které autor úmyslně použije velké množství nových slov v závěrečné části (např. v souvislosti s odhalením okolností zločinu); použít pro stanovení slovního bohatství počátečních sto či tisíc slov je v takovém případě jistě zavádějící. Není tedy překvapivé, že naprostá většina studií zabývajících se měřením slovního bohatství je proto zaměřena na hledání způsobů, jak vliv délky textu eliminovat, nejčastěji prostřednictvím nějaké transformace (srov. Baayen 1989; Bernet 1988; Covington – McFall 2010; Ejiri – Smith 1993; Guiraud 1954, 1959; Herdan 1960, 1966; Hess – Sefton – Landry 1986, 1989; Honoré 1979; Martynenko 2010; Menard 1983;

Müller 2002; Panas 2001; Popescu a kol.[i] 2009; Popescu – Čech – Altmann 2011c, 2012b; Ratkowsky, Hantrais 1975; Těšitelová 1972; Tuldava, 1995; Tuzzi – Popescu – Altmann 2010; Tweedie – Baayen 1998; Weitzman 1971; Yule 1944).

V následujících řádcích bude představeno pět metod, jak měřit slovní bohatství textu. Všechny jsou také aplikovány na analýzu básní E. Bachletové a následně jsou sledovány korelace mezi jednotlivými indexy vyjadřujícími slovní bohatství (srov. část 4.7). Vycházíme z předpokladu, že vzhledem k tomu, že neexistuje žádné nezávislé kritérium, na jehož základě by bylo možné posoudit kvalitu daného indexu, je analýza korelace mezi indexy dobrým způsobem, jak sledovat vhodnost či nevhodnost jednotlivých indexů.

V následujících kapitolách budou představeny tyto indikátory:

- index opakování,
- entropie,
- indikátor R_1, který je založen na distribuční funkci rankového rozdělení a na tzv. *h*-bodu,
- délka křivky, vyjadřující vztah mezi pořadím a frekvencí,
- Giniho koeficient.

4.1 INDEX OPAKOVÁNÍ SLOV

Jednou z možností, jak měřit slovní bohatství textu, je výpočet indexu opakování slov *RR* (repeat rate), který vyjadřuje míru koncentrovanosti textu vzhledem k použitému lexiku: čím je hodnota *RR* vyšší, tím menší je „rozptýlenost" slovníku, a tudíž i menší slovní bohatství textu (srov. Popescu a kol.[i] 2009). Index opakování slov je definován jako

$$RR = \sum_{r=1}^{V} p_r^2 \ , \tag{4.1}$$

kde V je velikost slovníku, tj. počet různých slovních tvarů[8] (types) v textu, a p_r pravděpodobnost výskytu slova *r*. Pokud tyto pravděpodobnosti určíme na základě relativních frekvencí slov v textu, tj.

8 Samozřejmě je možné pracovat také s lemmaty.

$$p_r = \frac{f_r}{N} \,, \tag{4.2}$$

kde f_r je absolutní frekvence daného slova a N celkový počet slov (tokens) v textu, můžeme vzorec (4.1) psát jako

$$RR = \frac{1}{N^2} \sum_{r=1}^{V} f_r^2 \,. \tag{4.3}$$

Pro ilustraci teoretické maximální a minimální hodnoty indexu opakování *RR* vyjděme z předpokladu, že máme text o délce $N = 100$ slov. V případě maximálně koncentrovaného slovníku by se celý takový text skládal pouze z jediného slova, které by se stokrát zopakovalo. V tomto případě by platilo

$$RR = \frac{100^2}{100^2} = 1 \,,$$

což je hodnota maximálně „chudého" slovníku. Na druhou stranu, pokud by se každé slovo v textu vyskytlo pouze jednou, tj. index opakování slov by byl nejmenší a slovník nejbohatší, pro nejnižší teoretickou hodnotu *RR* v textu o délce $N = 100$ slov platí

$$RR = \frac{1^2 + 1^2 + 1^2 + \cdots + 1^2}{100^2} = \frac{100}{100^2} = 0{,}01 \,.$$

Jak vidíme, minimální hodnota *RR* je závislá na velikosti slovníku, platí tedy

$$RR_{min} = \frac{1}{N^2} \sum_{r=1}^{V} \left(\frac{N}{V} \right)^2 = \frac{1}{V} \,, \tag{4.4}$$

což znamená, že hodnota *RR* leží v intervalu $< \frac{1}{V} ; 1 >$.

Pokud chceme testovat rozdíly *RR* mezi jednotlivými texty, musíme znát rozptyl *RR*. Hodnotu rozptylu $Var(RR)$, nutnou pro možnost porovnání textů, vypočítáme podle vzorce

$$Var(RR) = \frac{4}{N} \left(\sum_{r=1}^{V} p_r^3 - RR^2 \right) \,. \tag{4.5}$$

Pro porovnání textů použijeme opět asymptotický *u*-test,

$$u = \frac{|RR_1 - RR_2|}{\sqrt{Var(RR_1) + Var(RR_2)}} \,. \tag{4.6}$$

Pro ilustraci sledujme postup výpočtu *RR* u *Kazatele* (3,1–8) a u básně B. Trojaka *Září pod Čantoryjí* (viz Tab. 3.1 a 3.2 v kap. 3.1). Hodnoty *RR* u obou básní vypočítáme podle vzorce (4.3)

$$RR_{Kazatel} = \frac{30^2 + 14^2 + 7^2 + 2^2 + 2^2 + 2^2 + 2^2 + 1^2 + \cdots + 1^2}{96^2} =$$
$$= \frac{1198}{9216} = 0{,}129991 \ ,$$

$$RR_{Září\ pod\ Čantoryjí} = \frac{3^2 + 3^2 + 2^2 + 2^2 + 1^2 + \cdots + 1^2}{74^2} =$$
$$= \frac{90}{5476} = 0{,}016435 \ .$$

Rozptyl vypočítáme podle vzorce (4.5)

$$Var(RR_{Kazatel}) =$$
$$= \frac{4}{96}\left(\frac{30^3 + 14^3 + 7^3 + 2^3 + 2^3 + 2^3 + 2^3 + 1^3 + \cdots + 1^3}{96^3} - 0{,}129991^2\right) =$$
$$= \frac{4}{96}\,(0{,}034085 - 0{,}016898) = \frac{0{,}068748}{96} = 0{,}00071613 \ ,$$

$$Var\left(RR_{Září\ pod\ Čantoryjí}\right) =$$
$$= \frac{4}{74}\left(\frac{3^3 + 3^3 + 2^3 + 2^3 + 1^3 + \cdots + 1^3}{74^3} - 0{,}016435^2\right) =$$
$$= \frac{4}{74}\,(0{,}000331 - 0{,}00027) = \frac{0{,}000243}{74} = 0{,}00000327 \ .$$

Pro porovnání hodnot *RR* obou básní použijeme test (4.6)

$$u = \frac{|0{,}129991 - 0{,}016435|}{\sqrt{0{,}00071613 + 0{,}00000327}} = 4{,}23 \ ,$$

což znamená signifikantní rozdíl.

Pro snazší porovnání s ostatními indexy lze *RR* relativizovat, v našem případě použijeme relativizaci navrženou McIntoshem (1967), tj.

$$(4.7) \qquad RR_{MC} = \frac{1 - \sqrt{RR}}{1 - 1/\sqrt{V}} \ .$$

Pro *Kazatele* (3,1–8) a báseň B. Trojaka *Září pod Čantoryjí* dostáváme hodnoty

$$RR_{MC(Kazatel)} = \frac{1 - \sqrt{0{,}129991}}{1 - 1/\sqrt{44}} = 0{,}752972 \ ,$$

$$RR_{MC(Září\ pod\ Čantoryjí)} = \frac{1 - \sqrt{0{,}016435}}{1 - 1/\sqrt{68}} = 0{,}992112 \ .$$

4.2 ENTROPIE

Obecně se entropie pojímá jako míra neurčitosti systému. V případě analýzy frekvenční distribuce slov v textu chápeme entropii jako hodnotu vyjadřující míru diverzity – čím je hodnota entropie větší, tím diverzifikovanější (tj. méně koncentrovaný) je slovník, tudíž vysoká hodnota entropie je znakem velkého slovního bohatství. Pro výpočet entropie textu použijeme Shannonovu entropii definovanou jako

$$H = -\sum_{r=1}^{V} p_r \log_2 p_r \ , \tag{4.8}$$

kde p_r je relativní frekvence slov, definovaná v (4.2), přičemž platí, že $0 \log 0 = 0$. Tudíž ve výsledku dostáváme

$$H = \log_2 N - \frac{1}{N} \sum_{r=1}^{V} f_r \log_2 f_r \ . \tag{4.9}$$

Rozptyl *Var(H)*, nutný pro použití statistických testů, vypočítáme podle vzorce

$$\begin{aligned} Var(H) &= \frac{1}{N} \left(\sum_{r=1}^{V} p_r (\log_2 p_r)^2 - H^2 \right) = \\ &= \frac{1}{N} \left(\sum_{r=1}^{V} \frac{f_r}{N} \left(\log_2 \left(\frac{f_r}{N} \right) \right)^2 - H^2 \right) \ , \end{aligned} \tag{4.10}$$

takže je možné analogicky ke vzorci (4.6) testovat hodnotu entropie pomocí asymptotického *u*-testu, tj.

$$u = \frac{|H_1 - H_2|}{\sqrt{Var(H_1) + Var(H_2)}} \ . \tag{4.11}$$

Pro ilustraci teoretické maximální a minimální hodnoty entropie opět vyjděme z předpokladu, že máme text o délce $N = 100$ slov. V případě maximálně koncentrovaného slovníku by se celý takový text skládal pouze z jediného slova, které by se stokrát zopakovalo. V tomto případě by podle vzorce (4.9) platilo

$$H_{min} = \log_2 100 - \frac{100(\log_2 100)}{100} = \log_2 100 - \log_2 100 = 0 \quad .$$

Pokud by se každé slovo v textu vyskytlo pouze jednou, tj. index opakování slov by byl nejmenší a slovník nejbohatší, pro nejvyšší teoretickou hodnotu v textu o délce $N = 100$ slov platí

$$H_{max} = \log_2 100 - \frac{100 \sum_{r=1}^{V} \log_2 1}{100} = \log_2 100 = 6{,}64 \quad .$$

Vzhledem k tomu, že v textu s maximální entropií se každé slovo vyskytne jednou, platí $N = V$, tudíž můžeme maximální hodnotu entropie vyjádřit obecně takto

$$H_{max} = -\sum_{i=1}^{V} \frac{1}{V} \log_2 \frac{1}{V} = \log_2 V \quad . \tag{4.12}$$

Relativní hodnotu entropie ležící v intervalu <0, 1> získáme jednoduchým vydělením pozorované hodnoty entropie H její teoretickou maximální hodnotou H_{max}, tj.

$$H_{rel} = \frac{H}{H_{max}} = \frac{H}{\log_2 V} \quad . \tag{4.13}$$

Celý postup výpočtu a porovnání entropie budeme demonstrovat opět na textu *Kazatele* (3,1–8) a na básni B. Trojaka *Září pod Čantoryjí* (viz Tab. 3.1 a 3.2 v kap. 3.1). Hodnoty H v obou případech vypočítáme podle vzorce (4.9)

$$H_{Kazatel} = \log_2 96 - \frac{30 \cdot \log_2 30 + 14 \cdot \log_2 14 + \cdots + 1 \cdot \log_2 1}{96} =$$
$$= \log_2 96 - \frac{228{,}1612}{96} = 6{,}585 - 2{,}3767 = 4{,}2083 \ ,$$

$$H_{Září\ pod\ Čantoryjí} = \log_2 74 - \frac{3 \cdot \log_2 3 + 3 \cdot \log_2 3 + \cdots + 1 \cdot \log_2 1}{74} =$$
$$= \log_2 74 - \frac{13{,}5098}{74} = 6{,}2095 - 0{,}1826 = 6{,}0269 \ .$$

Maximální hodnota entropie je podle vzorce (4.12)

$$H_{max(Kazatel)} = \log_2 44 = 5{,}4594 \ ,$$

$$H_{max(Září\ pod\ Čantoryjí)} = \log_2 68 = 6{,}0875 \ ,$$

takže relativní hodnota entropie je podle vzorce (4.13)

$$H_{rel(Kazatel)} = \frac{4{,}2083}{5{,}4594} = \ 0{,}7708 \ ,$$

$$H_{rel(Září\ pod\ Čantoryjí)} = \frac{6{,}0269}{6{,}0875} = \ 0{,}9900 \ .$$

Pro aplikaci statistického testu je nutné vypočítat rozptyl, na základě vzorce (4.10) dostáváme

$$Var(H_{Kazatel}) =$$
$$= \frac{1}{96}\left[\left(\frac{30}{96}\left(\log_2 \frac{30}{96}\right)^2 + \frac{14}{96}\left(\log_2 \frac{14}{96}\right)^2 + \frac{7}{96}\left(\log_2 \frac{7}{96}\right)^2 + \right.\right.$$
$$\left.\left. + \frac{2}{96}\left(\log_2 \frac{2}{96}\right)^2 + \cdots + \frac{1}{96}\left(\log_2 \frac{30}{96}\right)^2\right) - 4{,}2083^2\right] = 0{,}04811 \ ,$$

$$Var(H_{Září\ pod\ Čantoryjí}) =$$
$$= \frac{1}{74}\left[\left(\frac{3}{74}\left(\log_2 \frac{3}{74}\right)^2 + \frac{3}{74}\left(\log_2 \frac{3}{74}\right)^2 + \frac{2}{74}\left(\log_2 \frac{2}{74}\right)^2 + \right.\right.$$
$$\left.\left. + \frac{2}{74}\left(\log_2 \frac{2}{74}\right)^2 + \cdots + \frac{1}{74}\left(\log_2 \frac{1}{74}\right)^2\right) - 6{,}0269^2\right] = 0{,}003031 \ .$$

Nyní je možné testovat rozdíly *H* podle vzorce (4.11)

$$u = \frac{|4{,}2083 - 6{,}0269|}{\sqrt{0{,}04811 + 0{,}003031}} = 8{,}02 \ ,$$

což znamená, že jde o signifikantní rozdíl.

4.3 INDEX R_1

Měření slovního bohatství prostřednictvím indexu R_1 je založeno na vlastnostech frekvenční struktury textu, konkrétně vychází z vlastností tzv. *h*-bodu (viz

kap. 3.1) a normalizovaných[9] kumulativních frekvencí slov, jejichž pořadí je vyšší než hodnota *h*-bodu. Konkrétně

$$R_1 = 1 - \left(F(h) - \frac{h^2}{2N} \right) = 1 - \left(\frac{\sum_{r=1}^{h} f_r}{N} - \frac{h^2}{2N} \right) , \tag{4.14}$$

kde $F(h)$ je kumulativní relativní frekvence nad *h*-bodem. Kumulativní distribuce nad *h*-bodem se odečítá od 1, berou se tedy do úvahy především slova s malou frekvencí – na Obr. 3.1 jde o tzv. oblast autosémantik (nikoliv však jenom o hapax legomena). Rozptyl $Var(R_1)$, nutný pro použití statistických testů, vypočítáme podle vzorce

$$Var(R_1) = \frac{F(h)\,[1 - F(h)]}{N} , \tag{4.15}$$

protože $F(h)$ není nic jiného než proporce; pro jednoduchost vynecháme kovarianci, která při porovnávání textů nehraje velkou roli. Pro porovnání textů použijeme opět asymptotický *u*-test

$$u = \frac{\left| R_{1(1)} - R_{1(2)} \right|}{\sqrt{Var(R_{1(1)}) + Var(R_{1(2)})}} . \tag{4.16}$$

Pro ilustraci uvedeme výpočet indexu R_1 opět u *Kazatele* (3,1–8) a u básně B. Trojaka *Září pod Čantoryjí* (viz Tab. 3.1 a 3.2 v kap. 3.1)

$$R_{1(Kazatel)} = 1 - \left(\frac{30 + 14 + 7}{96} - \frac{3{,}67^2}{2\,(96)} \right) = 1 - (0{,}5313 - 0{,}0702) = \\ = 0{,}5389 ,$$

$$R_{1(Září\ pod\ Čantoryjí)} = 1 - \left(\frac{3 + 3}{74} - \frac{2{,}5^2}{2\,(74)} \right) = 1 - (0{,}0812 - 0{,}0422) = \\ = 0{,}9611 .$$

Podle vzorce (4.15) vypočítáme rozptyl

$$Var\left(R_{1(Kazatel)}\right) = \frac{0{,}5313\,(1 - 0{,}5313)}{96} = 0{,}00259399 ,$$

9 Pro podrobné vysvětlení normalizace viz Popescu a kol.[i] (2009: 30) a Popescu – Čech – Altmann (2012b: 192).

$$Var\left(R_{1(Září\ pod\ Čantoryjí)}\right) = \frac{0{,}0812\,(1-0{,}0812)}{74} = 0{,}00100685\ ,$$

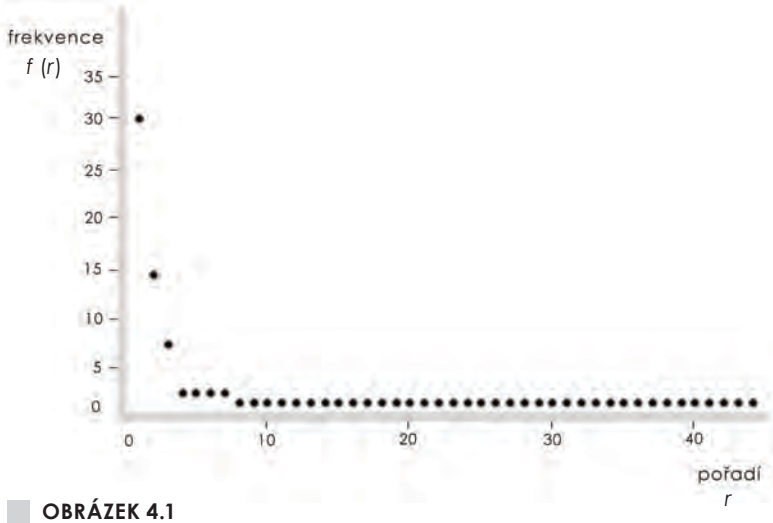
následně porovnáme indexy slovního bohatství R_1 u obou textů

$$u = \frac{|0{,}5389 - 0{,}9611|}{\sqrt{0{,}00259399 + 0{,}00100685}} = 7{,}04\ ,$$

což opět znamená signifikantní rozdíl.

4.4 DÉLKA KŘIVKY (*R* INDEX)

Měření slovního bohatství představené v této kapitole vychází z vlastností křivky vytvořené na základě grafického znázornění frekvence slov seřazených podle pořadí určeného právě frekvencí (srov. Popescu – Mačutek – Altmann 2009, kap. 5). Konkrétně, máme-li graf, na jehož ose *x* je číselná hodnota pořadí slov *r* a na ose *y* absolutní frekvence *f* těchto slov uspořádaných podle klesající frekvence, získáme soustavu bodů vyjadřujících vztah mezi těmito dvěma hodnotami; např. pro *Kazatele* (3,1–8) získáváme graf znázorněný na Obr. 4.1.

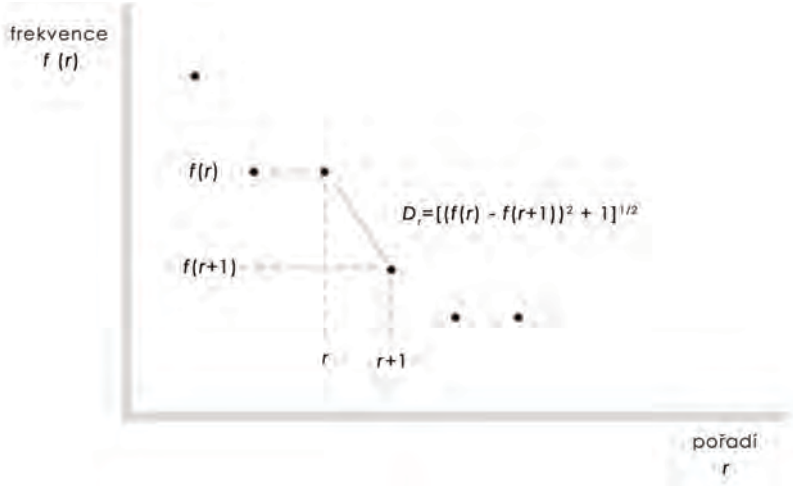


OBRÁZEK 4.1
Grafické vyjádření vztahu pořadí slova a jeho frekvence u *Kazatele* (3,1-8).

Pokud body spojíme, můžeme měřit vzdálenost mezi jednotlivými po sobě jdoucími body jako tzv. euklidovskou vzdálenost D_r, která je definována jako

(4.17) $$D_r = \sqrt{[f(r) - f(r+1)]^2 + 1}\ ,$$

srov. Obr. 4.2.



OBRÁZEK 4.2
Měření euklidovské vzdálenosti.

Délka celé křivky L je pak dána součtem jednotlivých vzdáleností, tj.

(4.18) $$L = \sum_{r=1}^{V-1} D_r = \sum_{r=1}^{V-1} \sqrt{[f(r) - f(r+1)]^2 + 1}\ .$$

Samotný index slovního bohatství R je definován vzhledem k h-bodu (viz kapitola 3.1), konkrétně je počítána délka křivky L_h nad h-bodem

(4.19) $$L_h = \sum_{r=1}^{h} \sqrt{[f\,(r) - f\,(r+1)]^2 + 1}$$

a její proporce vůči celkové délce L, tj.

$$R = 1 - \frac{L_h}{L} \,. \tag{4.20}$$

V případě, že hodnota *h* není celé číslo, tj. $r \neq f(r)$, je při výpočtu třeba vzít v potaz také délku křivky nad *h*-bodem, proto

$$L_h = \sum_{r=1}^{[h]-1} \sqrt{[f(r) - f(r+1)]^2 + 1} + \sqrt{[h - f([h])]^2 + (h - [h])^2} \,, \tag{4.21}$$

kde $[h]$ je celočíselná část desetinného čísla vyjadřujícího hodnotu *h*-bodu.

Celý výpočet a porovnání opět demonstrujme na příkladu *Kazatele* (3,1–8) a básně B. Trojaka *Září pod Čantoryjí* (viz Tab. 3.1 a 3.2 v kap. 3.1)

$$\begin{aligned} L_{Kazatel} = &\sqrt{(30-14)^2+1} + \sqrt{(14-7)^2+1} + \sqrt{(7-2)^2+1} + \\ &+ \sqrt{(2-2)^2+1} + \sqrt{(2-2)^2+1} + \sqrt{(2-2)^2+1} + \cdots + \\ &+ \sqrt{(1-1)^2+1} = 68{,}6155 \,, \end{aligned}$$

$$\begin{aligned} L_{Září\ pod\ Čantoryjí} = &\sqrt{(3-3)^2+1} + \sqrt{(3-2)^2+1} + \sqrt{(2-2)^2+1} + \\ &+ \sqrt{(2-1)^2+1} + \sqrt{(1-1)^2+1} + \cdots + \\ &+ \sqrt{(1-1)^2+1} = 67{,}8284 \,. \end{aligned}$$

Vzhledem k tomu, že u obou básní platí, že $r \neq f(r)$ (pro *Kazatele* (3,1–8) $h = 3{,}67$, pro *Září pod Čantoryjí* $h = 2{,}5$), použijeme pro výpočet L_h vzorec (4.21), tj.

$$\begin{aligned} L_{h(Kazatel)} = &\sqrt{(30-14)^2+1} + \sqrt{(14-7)^2+1} + \\ &+ \sqrt{(3{,}67-7)^2 + (3{,}67-3)^2} = 26{,}499 \,, \end{aligned}$$

$$\begin{aligned} L_{h(Září\ pod\ Čantoryjí)} = &\sqrt{(3-3)^2+1} + \\ &+ \sqrt{(2{,}5-3)^2 + (2{,}5-2)^2} = 1{,}7071 \,. \end{aligned}$$

Hodnota indexu slovního bohatství *R* je

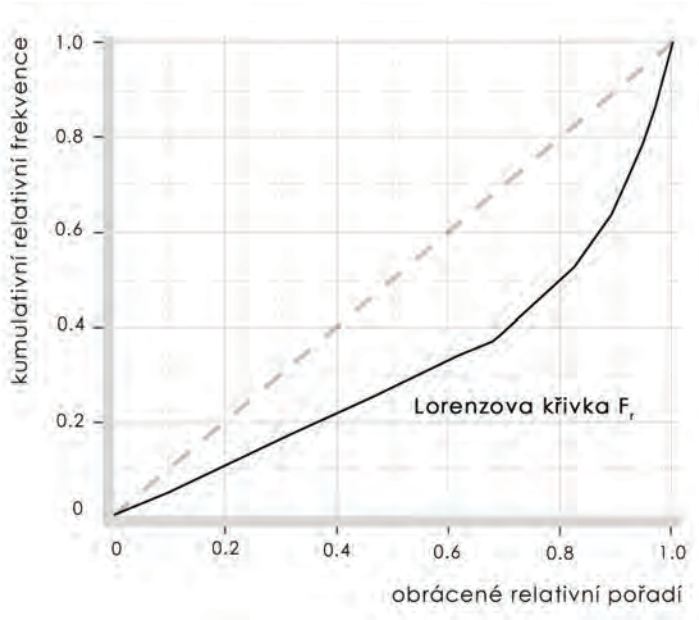
$$R_{Kazatel} = 1 - \frac{26{,}499}{68{,}6155} = 0{,}6138 \ ,$$

$$R_{Září\ pod\ Čantoryjí} = 1 - \frac{1{,}7071}{67{,}8284} = 0{,}9748 \ .$$

Výpočet rozptylu je relativně komplikovaný (srov. Popescu – Mačutek – Altmann 2010; Popescu – Altmann 2011), proto jej zde vzhledem k celkovému zaměření knihy (srov. *Úvod*) nebudeme uvádět. Obecně se ale dá říci, že po odvození rozptylu je pro porovnání jednotlivých textů možné použít *u*-test.

4.5 GINIHO KOEFICIENT

Výpočet Giniho koeficientu je možné provést pomocí Lorenzovy křivky, která graficky vyjadřuje kumulativní distribuční funkci slov v textu. Konkrétně vypočítáme-li z textu frekvenci slovních tvarů a seřadíme-li je kumulativně podle vzrůstající frekvence, získáváme grafické znázornění, které mimo jiné vyjadřuje i míru diverzifikace slovníku: sledujeme-li například grafické znázornění kumulativní relativní frekvenční distribuce na Obr. 4.3, vidíme, že 80 % různých slovních tvarů (types) (viz hodnota 0,8 na ose *x*) v daném textu reprezentuje 50 % všech výskytů slov (tokens) (viz hodnota 0,5 na ose *y*). Pokud by byla frekvenční distribuce slov naprosto pravidelná (tj. každé slovo by se vyskytovalo se stejnou frekvencí), pak by platilo $x = y$ (Lorenzova křivka by odpovídala přerušované čáře). Oblast mezi osou vyjadřující pravidelnou distribuci (tj. přerušovanou čárou) a Lorenzovou křivkou se nazývá Giniho koeficient. Je zřejmé, že čím je tato oblast větší, tím je míra slovního bohatství nižší (srov. Popescu a kol.[1] 2009: 57): pokud by se v textu vyskytoval pouze jediný slovní tvar (tj. slovník by byl maximálně chudý), byl by Giniho koeficient reprezentován pravoúhlým trojúhelníkem s vrcholy v bodech (0, 0; 1, 1; 1, 0), tj. jednalo by se o maximální možnou velikost plochy vyjádřenou tímto koeficientem.



OBRÁZEK 4.3
Lorenzova křivka vyjadřující kumulativní frekvenční distribuci.

Giniho koeficient (pro nenormalizované hodnoty) vypočítáme nejsnadněji (více viz Popescu a kol.[i] 2009, kap. 3.4) podle vzorce

$$G = \frac{1}{V}\left(V + 1 - \frac{2}{N}\sum_{r=1}^{V} r f(r)\right) = \frac{1}{V}\left(V + 1 - 2m_1'\right) \quad , \tag{4.22}$$

kde V je počet různých slovních tvarů v textu (types), N je počet všech slov (tokens), r je pořadí a $f(r)$ je frekvence slova v daném pořadí, m'_1 je průměr frekvenční distribuce, tj.

$$m_1' = \frac{\sum_{r=1}^{V} r f(r)}{N} \quad . \tag{4.23}$$

Pro přehlednost se hodnota slovního bohatství vypočítaná na základě Giniho koeficientu, označovaná většinou symbolem R_4, uvádí jako komplementární hodnota G, tj.

(4.24) $R_4 = 1 - G$.

Rozptyl Giniho koeficientu $Var(G)$ vypočítáme podle vzorce

(4.25) $Var(G) = \frac{4}{V^2}\ Var\left(m_1'\right) = \frac{4m_2}{V^2 N}$,

kde m_2 je rozptyl frekvenční distribuce, tj.

(4.26) $m_2 = \frac{1}{N}\sum_{i=1}^{V}\left(r_i - m_1'\right)^2 f(r_i)$.

Pro porovnání textů použijeme asymptotický *u*-test

(4.27) $u = \frac{|G_1 - G_2|}{\sqrt{Var(G_1) + Var(G_2)}}$.

Postup výpočtu Giniho koeficientu u *Kazatele* (3,1–8) a básně B. Trojaka *Září pod Čantoryjí* (viz Tab. 3.1 a 3.2 v kap. 3.1) je následující: nejdříve podle vzorce (4.23) vypočítáme průměr frekvenční distribuce m'_1

$$m'_{1(Kazatel)} = \frac{1\cdot 30 + 2\cdot 14 + 3\cdot 7 + 4\cdot 2 + 5\cdot 2 + \cdots + 44\cdot 1}{96} = 11{,}3021 ,$$

$$m'_{1(Září\ pod\ Čantoryjí)} = \frac{1\cdot 3 + 2\cdot 3 + 3\cdot 2 + 4\cdot 2 + 5\cdot 1 + \cdots + 68\cdot 1}{74} = 31{,}8783 .$$

Na základě vzorce (4.22) pak získáváme hodnoty Giniho koeficientu

$$G_{Kazatel} = \frac{1}{44}(44 + 1 - 2\cdot 11{,}3021) = 0{,}509 ,$$

$$G_{Září\ pod\ Čantoryjí} = \frac{1}{68}(68 + 1 - 2\cdot 31{,}8783) = 0{,}0771$$

a následně indexu R_4

$$R_{4(Kazatel)} = 1 - 0{,}509 = 0{,}491 ,$$

$$R_{4(Září\ pod\ Čantoryjí)} = 1 - 0{,}0771 = 0{,}9228 .$$

Výpočet rozptylu provedeme následujícím způsobem: rozptyl průměru frekvenční distribuce m_2 získáme na základě vzorce (4.26)

$$m_{2(Kazatel)} = \\ = \frac{(1-11{,}3021)^2 \cdot 30 + (2-11{,}3021)^2 \cdot 14 + \cdots + (44-11{,}3021)^2 \cdot 1}{96} = \\ = 180{,}9192 \ ,$$

$$m_{2(Září\ pod\ Čantoryjí)} = \\ = \frac{(1-31{,}8783)^2 \cdot 3 + (2-31{,}8783)^2 \cdot 3 + \cdots + (68-31{,}8783)^2 \cdot 1}{74} = \\ = 431{,}9987 \ ,$$

podle vzorce (4.25) pak určíme rozptyl *Var*(G)

$$Var(G_{Kazatel}) = \frac{4 \cdot 180{,}9192}{44^2 \cdot 96} = 0{,}003894 \quad ,$$

$$Var(G_{Září\ pod\ Čantoryjí}) = \frac{4 \cdot 431{,}9987}{68^2 \cdot 74} = 0{,}00505 \ .$$

Na základě vzorce (4.27) pak zjišťujeme, že

$$u = \frac{|0{,}509 - 0{,}0771|}{\sqrt{0{,}003894 + 0{,}00505}} = 4{,}57 \ ,$$

tzn. že se jedná o signifikantní rozdíl.

4.6 VZTAH INDEXŮ SLOVNÍHO BOHATSTVÍ A DÉLKY TEXTU

Vztah slovního bohatství a délky textu je na první pohled evidentní – čím je text delší, tím více různých slov zpravidla obsahuje, přičemž u dlouhých textů s jejich narůstající délkou pravděpodobnost výskytu nového slova klesá. Jak jsme uvedli v kapitole 4, většina indexů slovního bohatství se snaží vliv délky textu eliminovat prostřednictvím nějaké transformace, což je ostatně i případ všech indexů uvedených výše. Je však třeba zdůraznit, že žádný z nám známých indexů není na délce textu zcela nezávislý, a je otázkou, zda je takový index vzhledem k „přirozenému" vztahu mezi slovním bohatstvím a délkou vůbec stanovitelný. Jednou z možností, jak tento problém řešit, je například empirické odvození intervalu,

v němž je hodnota indexu na délce relativně nezávislá (srov. Čech 2014b). Obecně lze konstatovat, že je třeba při analýze slovního bohatství postupovat velmi obezřetně a mít stále na paměti možný vliv délky textu.

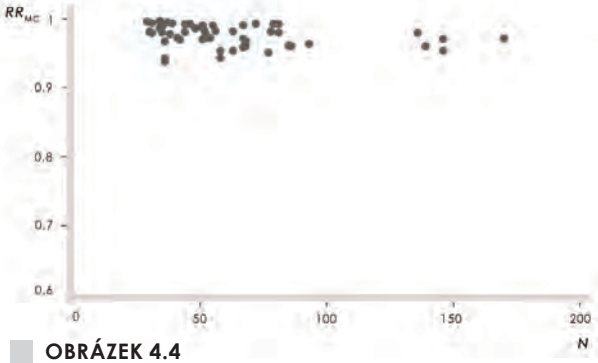
Pro ilustraci jsme se rozhodli aplikovat výše uvedené indexy na 52 básní E. Bachletové (srov. Tab. 4.1) a sledovat jejich vztah právě k délce textu; výsledky jsou prezentovány na Obr. 4.4–4.7

TABULKA 4.1
Hodnoty indexů měření slovního bohatství textu u 52 dvou básní E. Bachletové.

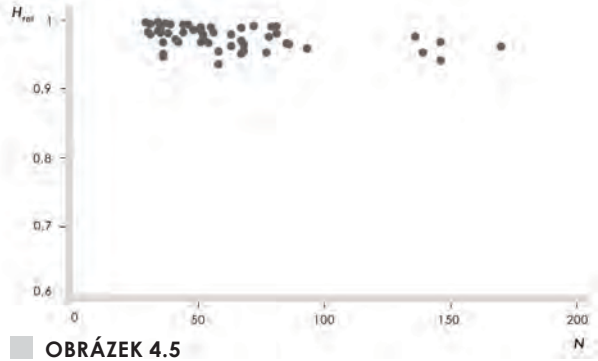
Báseň	***N***	***R***	R_1	RR_{mc}	H_{rel}	R_4
Miesto pre Nádej	29	1	0,9698	0,9963	0,9962	0,9667
Ťažko pokoriteľní	30	0,9452	0,9	0,9817	0,9818	0,8795
Tiché verše	31	0,9648	0,9355	0,9937	0,9933	0,9399
Ulomené zo slov	31	0,943	0,9032	0,9795	0,9782	0,84
Dovoľ mi slúžiť	34	1	0,9743	0,9971	0,9968	0,9715
Len áno	34	0,9621	0,9412	0,986	0,9834	0,8475
Bez rozlúčky	35	0,9682	0,9429	0,9925	0,9916	0,9223
Pravidlá odpúšťania	35	0,968	0,9063	0,9802	0,9805	0,8931
Tá Láska	35	0,966	0,9429	0,9889	0,9871	0,881
Dnešný luxus	36	0,9499	0,8924	0,9677	0,967	0,8075
Neopusť ma…	36	0,8665	0,8646	0,9384	0,9455	0,6637
Zbytočné srdce	36	0,8604	0,8333	0,9424	0,95	0,7798
Vďaka Pane!	37	0,9709	0,9459	0,9952	0,9945	0,949
Nado mnou Ty sám…	38	0,9355	0,8947	0,978	0,9799	0,8894
Vďaka za deň	39	0,9718	0,9487	0,9936	0,9926	0,9295

Báseň	***N***	***R***	R_1	RR_{mc}	H_{rel}	R_4
Istota	41	0,9575	0,9055	0,9727	0,9714	0,8271
Ešte raz	42	0,9274	0,869	0,9696	0,9674	0,811
Iba život	44	1	0,952	0,9924	0,9924	0,9556
Kým ich máme	44	0,9436	0,9091	0,981	0,9813	0,8928
Večerná ruža	46	1	0,965	0,9929	0,9928	0,9575
Čakáme šťastie…	48	0,9767	0,9401	0,9864	0,9851	0,9021
Spájania	48	0,9767	0,9401	0,9864	0,9851	0,9021
Do večnosti beží čas	51	0,94	0,8725	0,9706	0,9673	0,8083
Malé modlitby	51	0,9662	0,9412	0,9871	0,9838	0,8539
Precitnutie	51	0,9691	0,9412	0,9904	0,9887	0,9092
Vrátili sa	51	0,9691	0,9412	0,9904	0,9887	0,9092
Keď dohorí deň	52	0,9493	0,9062	0,972	0,9713	0,8408
Zasľúbenie jasu	52	0,9463	0,9231	0,9812	0,9778	0,8274
Ihly na nebi	54	0,9385	0,8981	0,9724	0,9661	0,773
Vyznania	55	0,971	0,9455	0,9905	0,9884	0,9006
Naše mamy	56	0,9713	0,9308	0,9827	0,981	0,8827
Som iná	58	0,931	0,8716	0,9534	0,9536	0,7715
To všetko je dar	58	0,9232	0,817	0,9433	0,935	0,7033
Aby spriesvitnela	63	0,9547	0,9127	0,9818	0,9783	0,855
Tak málo úsmevu	63	0,896	0,873	0,9537	0,9614	0,8516
Hľadanie odpovedí	67	0,9755	0,9552	0,9909	0,9878	0,8824

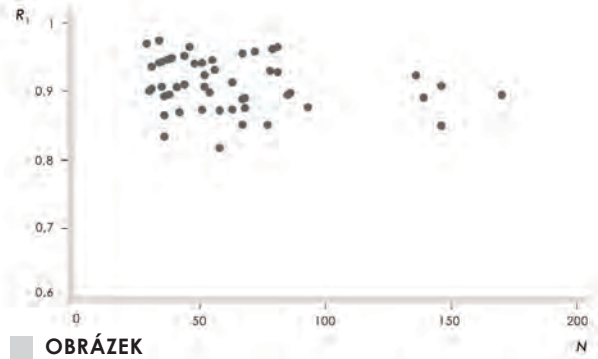
Báseň	***N***	***R***	R_1	RR_{mc}	H_{rel}	R_4
Naše svetlo	67	0,9284	0,8507	0,9589	0,9504	0,7396
Z neba do neba	67	0,927	0,8881	0,9709	0,9692	0,8339
Malý ošiaľ	68	0,8932	0,875	0,9607	0,9547	0,7301
Večerné ticho	68	0,9243	0,8897	0,9679	0,9638	0,8008
Idem za Tebou	72	0,9782	0,9583	0,9929	0,9905	0,9107
Čakanie na Boží jas	77	0,8972	0,8506	0,951	0,9521	0,7843
Rozťatá prítomnosť	78	0,9599	0,9295	0,9817	0,975	0,8056
Rozdelená bytosť	79	0,9797	0,962	0,9927	0,9897	0,8978
Čas pre nádych vône	81	0,9865	0,9645	0,9925	0,9902	0,9184
Prvotný sen	81	0,9578	0,9275	0,98	0,9798	0,9039
Podobnosť bytia	85	0,9087	0,8941	0,9613	0,9654	0,8541
Náš chrám	86	0,9209	0,8968	0,9607	0,9637	0,8446
Nepoznateľné	93	0,9253	0,8763	0,9636	0,9581	0,77
Dielo Stvoriteľa	136	0,9524	0,9228	0,9797	0,9751	0,8434
Iba neha	139	0,9194	0,8901	0,9603	0,9523	0,7243
Moje určenie	146	0,9301	0,9075	0,9707	0,9674	0,8104
Stály smútok pre šesť písmen	146	0,913	0,8493	0,9536	0,9407	0,6882
Vo večnosti slobodná	170	0,9544	0,8941	0,9716	0,9608	0,767



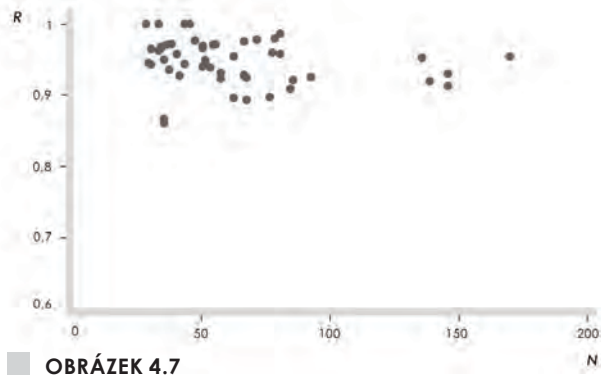
OBRÁZEK 4.4
Vztah indexu opakovaní RR_{MC} a délky textu *N* u 52 básní E. Bachletové.



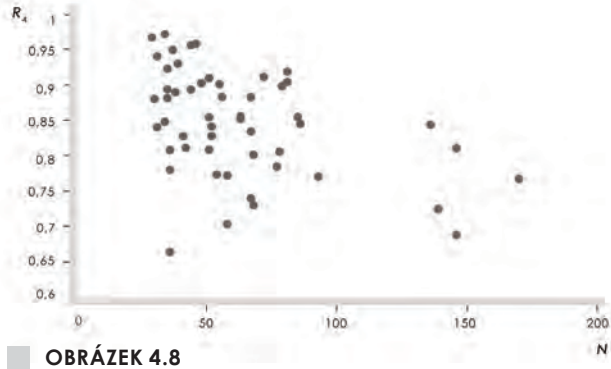
OBRÁZEK 4.5
Vztah relativní entropie H_{rel} a délky textu *N* u 52 básní E. Bachletové.



OBRÁZEK
Vztah indexu R_1 a délky textu *N* u 52 básní E. Bachletové.



OBRÁZEK 4.7
Vztah indexu *R* a délky textu *N* u 52 básní E. Bachletové.



OBRÁZEK 4.8
Vztah indexu slovního bohatství R_4 a délky textu *N* u 52 básní E. Bachletové.

S výjimkou indexu R_4 je i z pouhého grafického vyjádření zřetelné, že u analyzovaných textů, jejichž délka leží v intervalu $N \in \langle 29, 170 \rangle$, zjištěné hodnoty nevykazují závislost na délce textu. Dokonce i v případě indexu R_4, u něhož vzniká dojem, že jde o lineární pokles, koeficient determinace dává pro přímku hodnotu $R^2 = 0{,}1824$, což svědčí spíše o velkém rozptylu již v tomto malém intervalu než o nějaké závislosti.

4.7 KORELACE MEZI JEDNOTLIVÝMI INDEXY

Jak jsme uvedli v úvodu kapitoly týkající se metod měření slovního bohatství textu, vzhledem k tomu, že neexistuje žádné nezávislé kritérium, na jehož základě by bylo možné posoudit kvalitu daného indexu, je analýza korelace mezi jednotlivými indexy jedním ze způsobů, jak sledovat jejich vhodnost/nevhodnost. Tato analýza je cenná především proto, že jednotlivé indexy vycházejí z rozdílných charakteristik textu (srov. kap. 4.1–4.5).

Výsledky ukazují (viz Tab. 4.2), že všechny korelační koeficienty jsou signifikantní (52 stupňů volnosti), což se dá jednoduše ověřit *t*-testem.

TABULKA 4.2
Korelační koeficienty mezi indikátory slovního bohatství v básních E. Bachletové.

	R_1	RR_{Mc}	H_{rel}	R_4
R	0,88	0,91	0,86	0,79
R_1		0,94	0,94	0,84
RR_{Mc}			0,97	0,87
H_{rel}				0,94

Je také možné ukázat, že všechny indikátory jsou lineárními funkcemi ostatních indikátorů. Pro ilustraci uveďme vztah indexů R_1, RR_{Mc}, H_{rel} a R_4 k indikátoru R, kdy dostáváme

$$R_1 = -0{,}0769 + 1{,}0443R \qquad \text{s } R^2 = 0{,}77\,,$$
$$RR_{Mc} = 0{,}5512 + 0{,}4484R, \qquad \text{s } R^2 = 0{,}84\,,$$
$$H_{rel} = 0{,}5688 + 0{,}4277R, \qquad \text{s } R^2 = 0{,}74\,,$$
$$R_4 = -0{,}9144 + 1{,}8586R, \qquad \text{s } R^2 = 0{,}63\,,$$

přičemž všechny *t*-testy pro signifikanci parametrů a všechny *F*-testy pro signifikanci regrese dávají $P < 0{,}00001$. Tento výsledek znamená, že nezamítáme hypotézu o lineární závislosti, tedy jinými slovy náš předpoklad týkající vztahu závislosti mezi jednotlivými indexy není vyvrácen.

Z toho vyplývá, že všechny tyto indikátory detekují stejnou vlastnost, byť na základě jiných frekvenčních charakteristik textu. Na závěr jen připomeňme, že vztah mezi některými indikátory (např. *RR* a *H*) se dá vyjádřit čistě formálně (srov. Altmann 1988).

5 Míra aktivity a deskriptivity textu

Počátky měření deskriptivity (či ornamentality) a tzv. aktivity textu se pojí se jménem A. Busemanna (1925). Od té doby bylo navrženo několik modifikací původního postupu (srov. Altmann 1988; Antosch 1969; Bakker 1965; Boder 1940; Fischer 1969; Osgood – Walker 1959; Pieper 1979; Schlissmann 1948; Schubert 2008; Tuldava 2005; Wimmer – Altmann – Hřebíček – Ondrejovič – Wimmerová 2003). Obecně lze říct, že deskriptivita je v textu reprezentována adjektivy a aktivita slovesy (nejsou však většinou brána do úvahy stavová slovesa *být, mít, spát* atd.). Někteří autoři považují za nositele deskriptivity také adverbia, kterými se odpovídá na otázku „jak?". Dále je samozřejmě možné uvažovat také o verbálních substantivech (*nošení, bourání, rozbíjení* atd.) jako o výrazech vyjadřujících aktivitu. Je tedy evidentní, že existuje celá řada možných přístupů týkajících se definování deskriptivity a aktivity textu, přičemž vždy záleží na teoretických východiscích a cílech badatele. Důležité je však především to, že se deskriptivita (či aktivita) textu dá měřit pomocí kvantitativních metod, díky čemuž je možné použít tyto vlastnosti pro intersubjektivní porovnávání jednotlivých textů, žánrů, autorských stylů atd.

V této kapitole nejdříve představíme způsob měření aktivity (resp. deskriptivity) textu (5.1), dále se zaměříme na průběh vývoje této vlastnosti v textu (5.2) a na analýze lyrické poezie E. Bachletové ukážeme, jak lze texty vzhledem k tomuto způsobu měření klasifikovat (5.3, 5.4).

5.1 METODA MĚŘENÍ AKTIVITY A DESKRIPTIVITY TEXTU

Metoda měření aktivity a deskriptivity textu je velmi jednoduchá. Vyjdeme-li z předpokladu, že deskriptivita je v textu reprezentována množinou slov A obsahující adjektiva, adverbia, kterými se odpovídá na otázku „jak?", a nominalizovaná adjektiva, zatímco aktivita je reprezentovaná množinou slov V obsahující verba (kromě *být, mít* a modálních sloves *moci, smět, muset*) a deverbativní substantiva, pak celkovou aktivitu textu Q definujeme

$$Q = \frac{V}{V + A} \,. \tag{5.1}$$

Konkrétní postup ilustrujme na básni E. Bachletové *Z neba do neba* (symboly v textu *A* a *V* vyjadřují přítomnost výrazu reprezentujícího deskriptivitu či aktivitu).

Z neba do neba		*Obmýva*	V
Čistá	A	*Hojí*	V
Chladivá	A	*Napĺňa silou*	V
Upokojujúca	A	*Rozháňa staré*	V, A
Priezračná	A	*Skrútené*	A
Slobodná	A	*Triešti sa*	V
Chvejivá	A	*O skaly*	
Radostná	A	*A naše bolesti*	
Požehnaná	A	*Prebúdza pamäť*	V
		Osviežuje srdce	V
Voda		*Víri zabudnuté*	V, A
Chutí	V	*Navracia sa*	V
Kamením		*Prúdi*	V
Lístím		*Z času do času*	
Slnkom		*Z hôr do morí*	
Rozlieva sa	V	*Z morí do neba*	
Do nových organizmov	A	*Večný kolobeh*	A
Do nových vnemov	A	*Bytia*	
Do žíl		*Sveta*	
Človeka		*Božej rieky*	A
Duše		*života.*	

V básni se vyskytuje $A = 15$ slov vyjadřujících deskriptivitu a $V = 12$ slov vyjadřujících aktivitu, takže hodnota celkové aktivity básně je

$$Q = \frac{12}{12 + 15} = 0{,}4444 \ .$$

Na základě tohoto měření je samozřejmě možné porovnávat jednotlivé básně, autory, historické epochy, žánry atd. Konkrétně, rozdíly aktivity mezi jednotlivými texty můžeme testovat prostřednictvím normálního testu (srov. Altmann 1978), tj.

$$u = \frac{|Q_1 - Q_2|}{\sqrt{Q_1 Q_2}\sqrt{\frac{1}{V_1} + \frac{1}{A_1} + \frac{1}{A_2} + \frac{1}{V_2}}} \,. \tag{5.2}$$

Pokud je $u < 1{,}96$, přijímáme nulovou hypotézu, tj. nepokládáme změřený rozdíl za významný (na hladině významnosti $\alpha = 0{,}05$). Problém je v tom, že pro použití tohoto testu je třeba mít dostatečné počty adjektiv a verb v textu (což např. básně E. Bachletové nesplňují), protože jde o asymptotický test.

Druhou možností je testovat rozdíly s pomocí binomického rozdělení. Předpokládejme, že $n_1 Q_1$ jsou očekávané hodnoty jednoho textu, což znamená, že Q_1 je parametr p binomického rozdělení, přičemž $V_2 > V_1$, tj. v druhém textu se vyskytuje více V než v textu prvním. Nyní se ptáme, jaká je pravděpodobnost, že za těchto podmínek získáme $V \geq V_2$ ve druhém textu, který má n_2 (přičemž $n_x = V_x + A_x$). Počítáme tedy pravděpodobnost

$$P(X \geq V_2) = \sum_{x=V_2}^{n_2} \binom{n_2}{x} Q_1^x (1 - Q_1)^{n_2 - x} \,. \tag{5.3}$$

Konkrétně porovnáváme-li básně *Ihly na nebi* ($A = 6$, $V = 2$, $Q = 0{,}25$) a *Naše svetlo* ($A = 5$, $V = 8$, $Q = 0{,}62$), dosadíme hodnotu $Q_1 = 0{,}25$, $n_2 = 13$, dostáváme

$$P(X \geq 8) = \sum_{x=8}^{13} \binom{13}{x} 0{,}25^x (1 - 0{,}25)^{13 - x} = 0{,}0057 \,,$$

což znamená, že druhý text vykazuje signifikantně vyšší hodnotu aktivity než text první, protože vypočtená pravděpodobnost je menší než 0,05.

5.2 PRŮBĚH VÝVOJE AKTIVITY A DESKRIPTIVITY TEXTU

Jednou z otázek, které při úvahách o aktivitě a deskriptivitě textu vyvstávají, je otázka náhodnosti/nenáhodnosti průběhu vývoje těchto vlastností textu. Jinými slovy, je tento vývoj jen chaotickou oscilací pohybující se kolem průměru, nebo může být projevem jistých mechanismů souvisejících například s tématem textu, osobností autora, žánrem atd.?

Vývoj aktivity textu ve výše analyzované básni E. Bachletové *Z neba do neba* můžeme brát jako sekvenci hodnot vyjadřujících aktivitu, jak jsme si ji definovali výše, přičemž aktivitu budeme počítat kumulativně. V dané básni tedy máme následující řadu symbolů reprezentujících aktivitu (V) a deskriptivitu (A):

$$AAAAAAAAVVAAVVVVAAVVVVAVVAA\,.$$

Vzhledem k tomu, že v prvních osmi případech se jedná o slova vyjadřující deskriptivitu *A*, je ve všech těchto případech hodnota aktivity rovna nule. Devátá jednotka je symbol vyjadřující aktivitu *V*, v tomto případě počítáme hodnotu aktivity na základě vzorce (5.1) jako

$$Q = \frac{1}{1+8} = 0{,}1111\ .$$

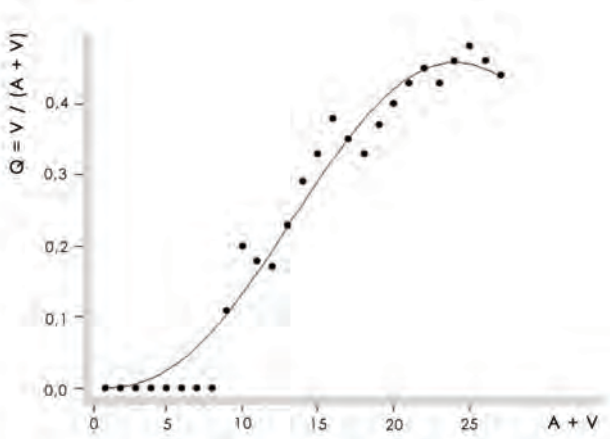
V dalším kroku počítáme se symbolem *V*, takže dostáváme

$$Q = \frac{2}{2+8} = 0{,}2\ .$$

Analogicky postupujeme celou sekvencí a výsledkem je následující kvantifikace průběhu aktivity básně E. Bachletové *Z neba do neba*:

0; 0; 0; 0; 0; 0; 0; 0; 0,1111; 0,2; 0,1818; 0,16667; 0,2308; 0,2857; 0,3333; 0,3750; 0,3528; 0,3333; 0,3684; 0,4; 0,4286; 0,4545; 0,4348; 0,4583; 0,48; 0,4613; 0,4444.

Již na první pohled je vidět (srov. Obr. 5.1), že se jedná o téměř monotónní nárůst aktivity pomalu směřující k rovnováze, která má hodnotu 0,5, což je případ, kdy jsou aktivita a deskriptivita vyrovnané. Tento průběh je samozřejmě možné modelovat prostřednictvím celé řady jednoduchých funkcí, my se ale pokusíme nejdříve najít teoretické zdůvodnění jejich volby.



OBR. 5.1
Průběh vývoje aktivity textu v básni *Z neba do neba*. Křivka vyjadřuje průběh beta funkce s parametry uvedenými v Tab. 5.1.

Vztah aktivity a deskriptivity textu je možné vnímat jako působení dvou protichůdných „sil", které se projevují při produkci textu. Proto použijeme Morseho funkci definovanou jako

$$y = a + b\left(1 - e^{-c(x-d)}\right)^2 , \tag{5.4}$$

jejíž derivací získáme

$$y' = 2bc\left(1 - e^{-c(x-d)}\right) e^{-c(x-d)} , \tag{5.5}$$

obecně tedy

$$y' = K f(x)\,[1 - f(x)] , \tag{5.6}$$

což můžeme interpretovat následovně: změna y' závisí na aktivní funkci $f(x)$, kterou vykonává mluvčí, přičemž je však „brzděn" komplementárním vlivem jazykové společnosti $1 - f(x)$.

Další možnost, jak modelovat působení aktivity a deskriptivity v textu, spočívá v přístupu, který je v kvantitativní lingvistice také běžně používán: relativní poměr změny (y'/y) je brán jako aditivní interakce dvou sil, tj.

$$\frac{y'}{y} = \frac{a}{x} - \frac{b}{M - x} \quad (5.7)$$

kde *a* je podíl mluvčího a *b* podíl posluchače, tj. síla, kterou působí na změnu nebo na konstrukci textu. Řešením této diferenciální rovnice je beta funkce, v našem případě definována jako

$$y = Cx^a(M - x)^b \quad (5.8)$$

kde *C* je integrační konstanta. Parametr *M* je shodný s maximální hodnotou *x*, ale v našem případě musí být menší než empirické maximum; dá se také vypočítat iteračně.

V Tab. 5.1 jsou uvedeny sekvence *Q* hodnot vyjadřujících průběh aktivity (kumulativní) v básních E. Bachletové a hodnoty parametrů Morseho funkce a beta funkce včetně koeficientu determinace. Tučně je označena hodnota *Q* celé básně a básně jsou seřazeny ve vzestupném pořadí vzhledem k této hodnotě. Pro lepší přehlednost jsou parametry z rovnice (5.4) označeny následujícím způsobem: *a* = *P1*, *b* = *P2*, *c* = *P3* a *d* = *P4*.

TABULKA 5.1
Sekvence Q hodnot vyjadřujících průběh aktivity (kumulativní) v básních E. Bachletové doplněná o hodnoty parametrů Morseho funkce a beta funkce.

Báseň	
Naše dejiny	AAAAA (A = 5, V = 0); 0; 0; 0; 0; **0**; beta: C = 0; R^2 = 1 Morse: P1 = 0; P2 = 0; P3 = 0; P4 = 0; R^2 = 1
To všetko je dar	AAAAAAAA (A = 8, V = 0); 0; 0; 0; 0; 0; 0; 0; **0**; beta: C = 0; R^2 = 1 Morse: P1 = 0; P2 = 0; P3 = 0; P4 = 0; R^2 = 1
Každodennosť	AAA (A = 3, V = 0); 0; 0; **0**; beta: C = 0; R^2 = 1 Morse: P1 = 0; P2 = 0; P3 = 0; P4 = 0; R^2 = 1

Báseň	
Večerné ticho	AAAAAAAAVAV (A = 9, V = 2); 0; 0; 0; 0; 0; 0; 0; 0; 0,11; 0,10; **0,18**; beta: C = 5,7973E-010; a = 8,1458; b = 0,5799; M = 12; R^2 = 0,91 Morse: P1 = 0; P2 = 5,3187; P3 = 0,0147; P4 = 0; R^2 = 0,65
Ihly na nebi	AAAVVAAA (A = 6, V = 2); 0; 0; 0; 0,25; 0,40; 0,33; 0,29; **0,25**; beta: C = 2,5821E-015; a = 6,2491; b = 9,7537; M = 15; R^2 = 0,90 Morse: P1 = 0; P2 = 0,4031; P3 = 0,3007; P4 = 0; R^2 = 0,66
Otázka	AAAV (A = 3, V = 1); 0; 0; 0; **0,25**; beta: C = 8,8261E-014; a = 22,8543; b = -0,9889; M = 25; R^2 = 1 Morse: P1 = 0; P2 = 1259,0652; P3 = 0,0030; P4 = 0; R^2 = 0,63
Nemám rada bielu	AAAAAVAAAVVAVAVAAAVAAV (A = 15, V = 7); 0; 0; 0; 0; 0; 0,17; 0,14; 0,12; 0,11; 0,20; 0,27; 0,25; 0,31; 0,29; 0,33; 0,31; 0,29; 0,28; 0,32; 0,30; 0,29; **0,32**; beta: C = 9,7961E-8; a = 2,3546; b = 2,7272; M = 38; R^2 = 0,92 Morse: P1 = 0; P2 = 0,3788; P3 = 0,1286; P4 = 0; R^2 = 0,89
Zbytočné srdce	AAAVVAAVAAV (A = 7, V = 4); 0; 0; 0; 0,25; 0,40; 0,33; 0,29; 0,38; 0,33; 0,30; **0,36**; beta: C = 7,1195E-011; a = 2,7389; b = 5,8925; M = 25, R^2 = 0,81 Morse: P1 = 0; P2 = 0,3936; P3 = 0,3215; P4 = 0; R^2 = 0,75
Tak málo úsmevu	VAAAAAAVVVVAVAAAAAV (A = 12, V = 7); 1; 0,5; 0,33; 0,25; 0,20; 0,17; 0,14; 0,25; 0,33; 0,40; 0,45; 0,42; 0,46; 0,43; 0,40; 0,38; 0,35; 0,33; **0,37**; Morse: P1 = 0,2025; P2 = 0,2178; P3 = 0,2867; P4 = 4,7363; R^2 = 0,91

Báseň	
Iba v modlitbe	AAAVAVVA (A = 5, V = 3); 0; 0; 0; 0,25; 0,20; 0,33; 0,43; **0,38**; beta: C = 0,00012; a = 3,4505; b = 1,2884; M = 10; R^2 = 0,92 Morse: P1 = 0; P2 = 1,1893; P3 = 0,1164; P4 = 0; R^2 = 0,87
Prvotný sen	AVVVVAAAAAAAAVVVAAAAAVVA (A = 15, V = 9); 0; 0,50; 0,67; 0,75; 0,80; 0,67; 0,57; 0,50; 0,44; 0,40; 0,36; 0,33; 0,31; 0,36; 0,40; 0,44; 0,41; 0,39; 0,37; 0,35; 0,33; 0,36; 0,39; **0,38**; beta: C = 1,5408E-014; a = 0,4401; b = 6,6996; M = 100; R^2 = 0,33 Morse: P1 = 0; P2 = 0,4534; P3 = 1,2458; P4 = 0; R^2 = 0,19
Náš chrám	AAAVAAAVVAAAAVAVVVVAA (A = 13, V = 8); 0; 0; 0; 0,25; 0,20; 0,17; 0,14; 0,25; 0,33; 0,30; 0,27; 0,25; 0,23; 0,29; 0,27; 0,31; 0,35; 0,39; 0,42; 0,40; **0,38**; beta: C = 0,00012; a = 0,8919; b = 1,3310; M = 79; R^2 = 0,80 Morse: P1 = 0; P2 = 0,3718; P3 = 0,1958; P4 = 0; R^2 = 0,79
Dielo Stvoriteľa	VAAAAAAAAAAAAAVAVAVVVAVVAVVVVAAAVA; (A = 21, V = 13); 1; 0,50; 0,33; 0,25; 0,20; 0,17; 0,14; 0,13; 0,11; 0,10; 0,09; 0,08; 0,08; 0,07; 0,13; 0,13; 0,18; 0,17; 0,21; 0,25; 0,28; 0,27; 0,30; 0,33; 0,32; 0,35; 0,37; 0,39; 0,41; 0,40; 0,39; 0,38; 0,39; **0,38**; Morse: P1 = 0,0606; P2 = 0,3921; P3 = 0,1099; P4 = 9,1328; R^2 = 0,94
Večerná ruža	AAAVAVVVAAVA (A = 7, V = 5); 0; 0; 0; 0,25; 0,20; 0,33; 0,43; 0,50; 0,44; 0,40; 0,45; **0,42;** beta: C = 0,00002; a = 2,4961; b = 2,0973; M = 17,47; R^2 = 0,92 Morse: P1 = 0; P2 = 0,5484; P3 = 0,2308; P4 = 0; R^2 = 0,86

Báseň	
Tiché verše	VAVVAAA (A = 4, V = 3); 1; 0,50; 0,67; 0,75; 0,60; 0,50; **0,43**; Morse: P1 = 0,4104; P2 = 0,1778; P3 = 2,5173; P4 = 1,4121; R^2 = 0,70
Z neba do neba	AAAAAAAAVVAAVVVVAAVVVVAVVAA (A = 15, V = 12); 0; 0; 0; 0; 0; 0; 0; 0; 0,11; 0,20; 0,18; 0,17; 0,23; 0,29; 0,33; 0,38; 0,35; 0,33; 0,37; 0,40; 0,43; 0,45; 0,43; 0,46; 0,48; 0,46; **0,44**; beta: C = 0,00001; a = 2,4436; b = 1,1049; M = 35; R^2 = 0,96 Morse: P1 = 0; P2 = 0,7930; P3 = 0,0594; P4 = 0; R^2 = 0,94
Moje určenie	AVAVAVVAAVAVAAVAAAVAVVAVVVVAA (A = 15, V = 14); 0; 0,50; 0,33; 0,50; 0,40; 0,50; 0,57; 0,50; 0,44; 0,50; 0,45; 0,50; 0,46; 0,43; 0,47; 0,44; 0,41; 0,39; 0,42; 0,40; 0,43; 0,45; 0,43; 0,46; 0,48; 0,50; 0,52; 0,50; **0,48**; Morse: P1 = 0; P2 = 0,4641; P3 = 0,9071; P4 = 0; R^2 = 0,58
Podobnosť bytia	AVVVAVAVVAVVAAVAVAAA (A = 10, V = 10); 0; 0,50; 0,67; 0,75; 0,60; 0,67; 0,57; 0,63; 0,67; 0,60; 0,64; 0,67; 0,62; 0,57; 0,60; 0,56; 0,59; 0,56; 0,53; **0,50;** Morse: P1 = 0; P2 = 0,6142; P3 = 0,9338; P4 = 0; R^2 = 0,67
Zasľúbenie jasu	VVVVVAAAAA (A = 5, V = 5); 1; 1; 1; 1; 1; 0,83; 0,71; 0,63; 0,56; **0,50**; Morse: P1 = 0,0674; P2 = 2,1030; P3 = 0,0154; P4 = 35,7696; R^2 = 0,83

Báseň	
Čas pre nádych vône	AVVVAVVAAVVVAAAVAAVVAAAV (A = 12, V = 12); 0; 0,50; 0,67; 0,75; 0,60; 0,67; 0,71; 0,63; 0,56; 0,60; 0,63; 0,67; 0,62; 0,57; 0,53; 0,56; 0,53; 0,50; 0,53; 0,55; 0,52; 0,50; 0,48; **0,50;** Morse: P1 = 0; P2 = 0,5868; P3 = 0,9851; P4 = 0; R^2 = 0,56
Do večnosti beží čas	VAAVAVAAVAAVVV (A = 7, V = 7); 1; 0,5; 0,33; 0,50; 0,40; 0,50; 0,43; 0,38; 0,44; 0,40; 0,36; 0,42; 0,46; **0,50**; Morse: P1 = 0,4099; P2 = 0,0235; P3 = 0,8094; P4 = 3,2172; R^2 = 0,90
Iba život	VAAAVVVVVAAA (A = 6, V = 6); 1; 0,50; 0,33; 0,25; 0,40; 0,50; 0,57; 0,63; 0,67; 0,60; 0,55; **0,50**; Morse: P1 = 0,2500; P2 = 0,5259; P3 = 0,2662; P4 = 4,000; R^2 = 0,64
Spájanie	AAVVVVVAA (A = 4,V = 5); 0; 0; 0,33; 0,50; 0,60; 0,67; 0,71; 0,63; **0,56;** beta: C = 2,3941E-006; a = 2,7790; b = 3,4605; M = 15; R^2 = 0,96 Morse: P1 = 0; P2 = 0,7406; P3 = 0,3705; P4 = 0; R^2 = 0,86
Dnešný luxus	AVAVVVVAA (A = 4, V = 5); 0; 0,50; 0,33; 0,50; 0,60; 0,67; 0,71; 0,63; **0,56**; beta: C = 0,0557; a = 1,0404; b = 0,4417; M = 10; R^2 = 0,80 Morse: P1 = 0; P2 = 0,6503; P3 = 0,5959; P4 = 0; R^2 = 0,78
Istota	VAAAVAVVV (A = 4, V = 5); 1; 0,50; 0,33; 0,25; 0,40; 0,33; 0,43; 0,50; **0,56**; Morse: P1 = 0,2863; P2 = 0,4412; P3 = 0,2865; P4 = 03,8614 R^2 = 0,97

Báseň	
Smútok	AVVVAAV (A = 3, V = 4); 0; 0,50; 0,67; 0,75; 0,60; 0,50; **0,57**; beta: C = 0,00014; a 1,6831; b = 2,9697; M = 12; R^2 = 0,79 Morse: P1 = 0; P2 = 0,6368; P3 = 0,8804; P4 = 0; R^2 = 0,69
Keď dohorí deň	VVVVAAVVVVAVVAAVAAA (A = 8, V = 11); 1; 1; 1; 1; 0,80; 0,67; 0,71; 0,75; 0,78; 0,80; 0,73; 0,75; 0,77; 0,71; 0,67; 0,69; 0,65; 0,61; **0,58;** Morse: P1 = 0,7017; P2 = 0,0053; P3 = 0,1326; P4 = 17,2315; R^2 = 0,66
Nado mnou Ty sám	AVVVVVVAAAVA (A = 5, V = 7); 0; 0,50; 0,67; 0,75; 0,80; 0,83; 0,86; 0,75; 0,67; 0,60; 0,64; **0,58**; beta: C = 3,5377E-019; a = 1,4731; b = 10,4847; M = 50; R^2 = 0,89 Morse: P1 = 0; P2 = 0,7312; P3 = 0,8019; P4 = 0; R^2 = 0,73
Iba neha	VAAVAAAAVVVVVVVVAAAVVVVAAAAAAAVVVVVVVVV (A = 16, V = 23); 1; 0,50; 0,33; 0,50; 0,40; 0,33; 0,29; 0,25; 0,33; 0,40; 0,45; 0,50; 0,54; 0,57; 0,60; 0,62; 0,59; 0,56; 0,53; 0,55; 0,57; 0,59; 0,61; 0,58; 0,56; 0,54; 0,52; 0,50; 0,48; 0,47; 0,48; 0,50; 0,52; 0,53; 0,54; 0,56; 0,57; 0,58; **0,59**; Morse: P1 = 0,3129; P2 = 0,2408; P3 = 0,2614; P4 = 4,8712; R^2 = 0,75
Nepoznateľné	VVVVVVAAVAVAAVAVAVVVVAAVA (A = 10, V = 15); 1; 1; 1; 1; 1; 1; 0,86; 0,75; 0,78; 0,70; 0,73; 0,67; 0,62; 0,64; 0,60; 0,63; 0,59; 0,61; 0,63; 0,65; 0,67; 0,64; 0,61; 0,63; **0,60;** Morse: P1 = 1,0491; P2 = -0,4464; P3 = 0,2418; P4 = 2,4551; R^2 = 0,94

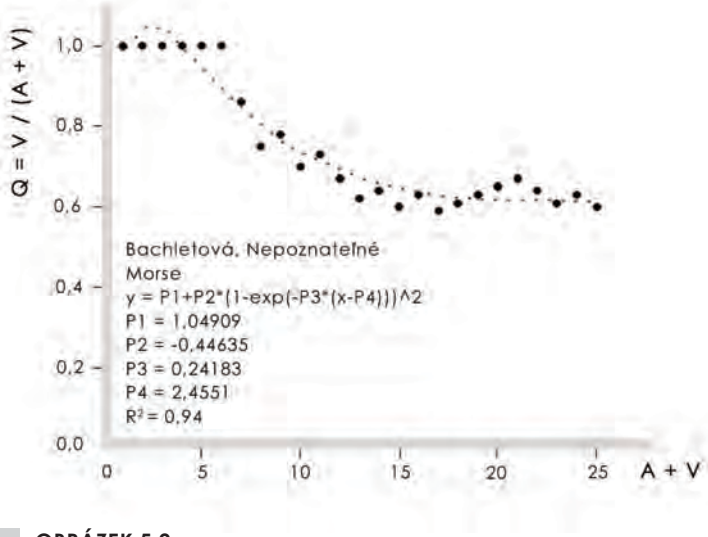
Báseň	
Návraty	AVVAV (A = 2, V = 3); 0; 0,50; 0,67; 0,50; **0,60**; beta: C = 0,0001; a = 2,1640; b = 3,3825; M = 9,46; R^2 = 0,79 Morse: P1 = 0; P2 = 0,6527; P3 = 0,7800; P4 = 0; R^2 = 0,73
Vyznania	AAAVAAAVVAVVVVVAVVVV (A = 8, V = 12); 0; 0; 0; 0,25; 0,20; 0,17; 0,14; 0,25; 0,33; 0,30; 0,36; 0,42; 0,46; 0,50; 0,53; 0,50; 0,53; 0,56; 0,58; **0,60**; Morse: P1 = 0; P2 = 0,7404; P3 = 0,1142; P4 = 0; R^2 = 0,94
Som iná	VVVAVAAVVVAAVV (A = 5, V = 8); 1; 1; 0,67; 0,75; 0,60; 0,50; 0,57; 0,63; 0,67; 0,60; 0,55; 0,58; **0,62;** Morse: P1 = 0,5817; P2 = 0,1106; P3 = 0,1517; P4 = 8,3665; R^2 = 0,82
Naše svetlo	VVVAAVVVAAAVV (A = 5; V = 8); 1; 1; 1; 0,75; 0,60; 0,67; 0,71; 0,75; 0,67; 0,60; 0,55; 0,58; **0,62**; Morse: P1 = 0,5616; P2 = 5,7977E-6; P3 = 0,1199; P4 = 48,3456; R^2 = 0,80
Precitnutie	AVAAVVVV (A = 3, V = 5); 0; 0,50; 0,33; 0,25; 0,40; 0,50; 0,57; **0,63**; Morse: P1 = 0; P2 = 0,5514; P3 = 0,5360; P4 = 0; R^2 = 0,56
Hľadanie odpovedí	VVAVVVAAVAVVAAVVVVAAVVV (A = 8, V = 15); 1; 1; 0,67; 0,75; 0,80; 0,83; 0,71; 0,63; 0,67; 0,60; 0,54; 0,67; 0,62; 0,57; 0,60; 0,63; 0,65; 0,67; 0,63; 0,60; 0,62; 0,64; **0,65;** Morse: P1 = 0,6245; P2 = 0,0008; P3 = 0,1524; P4 = 21,5976; R^2 = 0,75
Čakáme šťastie	VAVAVVVVA (A = 3, V = 6); 1; 0,50; 0,67; 0,50; 0,60; 0,67; 0,71; 0,75; **0,67;** Morse: P1 = 0,5250; P2 = 0,1804; P3 = 0,6449; P4 = 2,4892; R^2 = 0,81

Báseň	
Vrátili sa	VAAVVVAVVVVA (A = 4, V = 8); 1; 0,5; 0,33; 0,50; 0,60; 0,67; 0,63; 0,67; 0,70; 0,73; **0,67;** Morse: P1 = 0,4030; P2 = 0,3135; P3 = 0,5355; P4 = 2,6261; R^2 = 0,94
Mladé oči	VAAVVV (A = 2, V = 4); 1; 0,50; 0,33; 0,50; 0,60; **0,67;** Morse: P1 = 0,3864; P2 = 0,6111; P3 = 0,3820; P4 = 2,8229; R^2 = 0,98
Stály smútok pre šesť písmen	AVVAAAAVAVVAAVVAAVVVVVVVVVVVAVVVV (A = 11, V = 22); 0; 0,50; 0,67; 0,50; 0,40; 0,33; 0,29; 0,38; 0,33; 0,40; 0,45; 0,42; 0,38; 0,43; 0,47; 0,44; 0,41; 0,44; 0,47; 0,50; 0,52; 0,55; 0,57; 0,58; 0,60; 0,62; 0,63; 0,64; 0,62; 0,63; 0,65; 0,66; **0,67;** Morse: P1 = 0,3650; P2 = 0,4207; P3 = 0,0802; P4 = 7,8894; R^2 = 0,84
Neopusť ma	VVAVVAVAVV (A = 3, V = 7); 1; 1; 0,67; 0,75; 0,80; 0,67; 0,71; 0,63; 0,67; **0,70;** Morse: P1 = 0,6778; P2 = 0,0642; P3 = 0,1608; P4 = 8,4994; R^2 = 0,74
Čakanie na Boží jas	VVAAAAAVVVVVVVVV (A = 5 , V = 12) 1; 1; 0,67; 0,50; 0,40; 0,33; 0,29; 0,37; 0,44; 0,50; 0,55; 0,58; 0,62; 0,64; 0,67; 0,69; **0,71;** Morse: P1 = 0,3587; P2 = 0,7920; P3 = 0,1139; P4 =6,8979; R^2 = 0,89
Bez rozlúčky	VAAVVVV (A = 2, V = 5); 1; 0,50; 0,33; 0,50; 0,60; 0,67; **0,71;** Morse: P1 = 0,3864; P2 = 0,6111; P3 = 0,3820; P4 = 2,8229; R^2 = 0,98
Vo večnosti slobodná	VVVVVVAAVAAAVVVVVVVVVA (A = 6, V = 17); 1;1;1;1;1;1; 0,86; 0,75; 0,78; 0,70; 0,64; 0,58; 0,61; 0,64; 0,67; 0,69; 0,71; 0,72; 0,74; 0,75; 0,76; 0,79; **0,74;** Morse: P1 = 0,6744; P2 = 95,5368; P3 = 0,0046; P4 = 15,18390; R^2 = 0,80

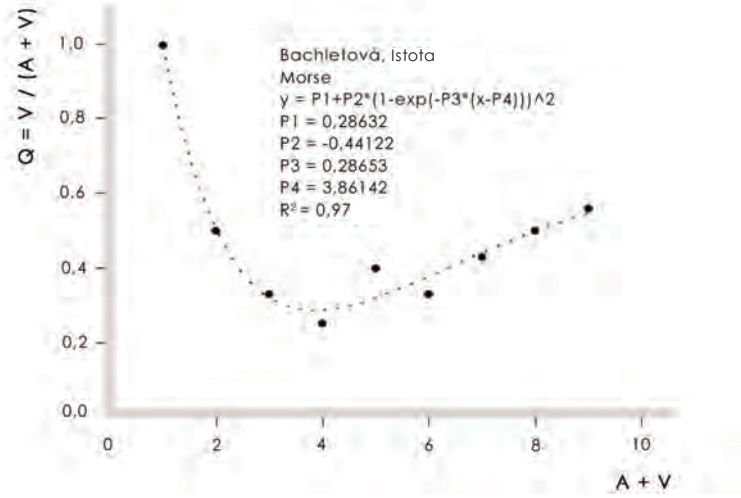
Báseň	
Neha domova	AVVV (A = 1, V = 3); 0; 0,50; 0,67; **0,75;** Beta: C = 0,3236; a = 0,0163; b = 0,3398; M = 10; R^2 = 0,95 Morse: P1 = 0; P2 = 1,06536; P3 = 0,49248; P4 = 0; R^2 = 0,90
Kým ich máme	VAVVVAVVVVAV (A = 3, V = 9); 1; 0,50; 0,67; 0,75; 0,80; 0,67; 0,71; 0,75; 0,78; 0,80; 0,73; **0,75;** Morse: P1 = 0,4184; P2 = 0,3318; P3 = 1,5202; P4 = 1,5548; R^2 = 0,90
Rozdelená bytosť	AAAVVAVVVVVVAVVVVVVV (A = 5, V = 15); 0; 0; 0; 0,25; 0,40; 0,33; 0,43; 0,50; 0,56; 0,60; 0,64; 0,67; 0,62; 0,64; 0,67; 0,69; 0,71; 0,72; 0,74; **0,75;** Morse: P1 = 0; P2 = 0,7687; P3 = 0,1980; P4 = 0; R^2 = 0,96
Malý ošiaľ	VVVVVAVVAVVVAV (A = 3, V = 11); 1; 1; 1; 1; 1; 0,83; 0,86; 0,88; 0,78; 0,80; 0,82; 0,83; 0,77; **0,79;** Morse: P1 = 0,7855; P2 = 0,5339; P3 = -0,0979; P4 = 12,5580; R^2 = 0,83
Ešte raz	VVAVV (A = 1, V = 4); 1; 1; 0,67; 0,75; **0,80**; Morse: P1 = 0,7465; P2 = 24,5245; P3 = 0,0358; P4 = 3,9126; R^2 = 0,67
Rozťatá prítomnosť	VVAVVVVVVVAVVVVVVVA (A = 3, V = 16); 1; 1; 0,67; 0,75 ;0,80; 0,83; 0,86; 0,88; 0,89; 0,90; 0,82; 0,83; 0,85; 0,86; 0,87; 0,88; 0,88; 0,89; **0,84** Morse: P1 = 0,7883; P2 = 0,0871; P3 = 0,3511; P4 = 3,8146; R^2 = 0,49

Báseň	
Malé modlitby	VVVVAVVVVVVV (A = 1, V = 11); 1; 1; 1; 1; 0,80; 0,83; 0,86; 0,88; 0,89; 0,90; 0,91; **0,92;** beta: C = 1,2084; a = -0,1071; b = -0,0600; M = 13; R^2 = 0,48 Morse: P1 = 0,8972; P2 = 0,0007; P3 = 0,4062; P4 = 7,6420; R^2 = 0,34
Zázrak	VV (A = 0, V = 2); 1; **1**; beta: C = 1,0; a = b = 0; R^2 = 1,0 Morse: P1 = 1; P2 = 0; P3 = 0; P4 = 0; R^2 = 1,0

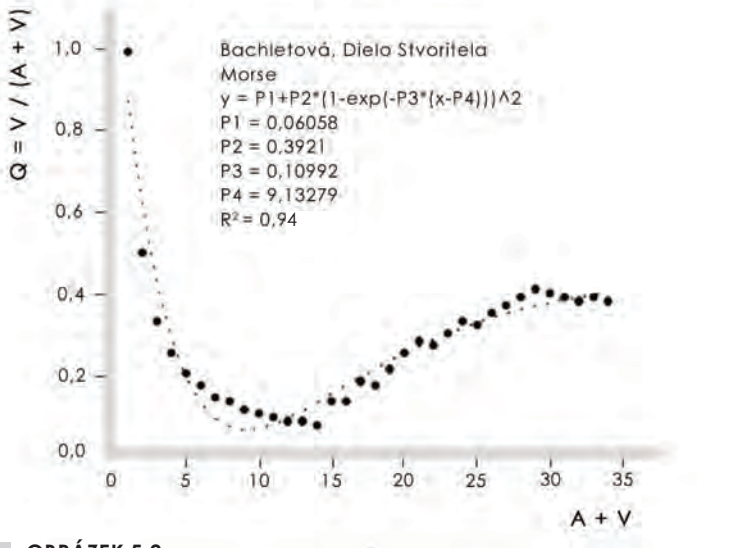
Z Tabulky 5.1 je patrné, že ne všechny básně lze dobře modelovat pomocí těchto dvou funkcí, což znamená, že je třeba dále rozpracovat teorii (a také samozřejmě testovat model na větším množství textů). Ukazuje se ale, že sekvence začínající nulou lze lépe modelovat prostřednictvím beta funkce, zatímco sekvence začínající hodnotou jedna prostřednictvím Morseho funkce. Na Obr. 5.2 můžeme sledovat průběh vývoje aktivity u vybraných básní E. Bachletové.



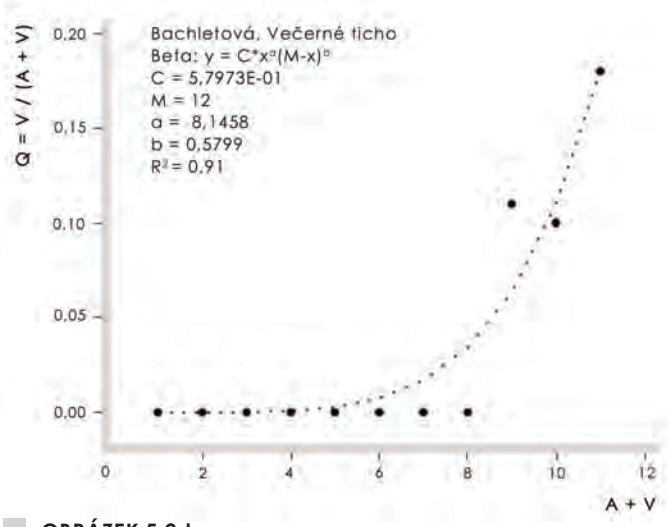
OBRÁZEK 5.2a
Průběh aktivity textu modelovaný prostřednictvím Morseho funkce v básni E. Bachletové *Nepoznateľné*.



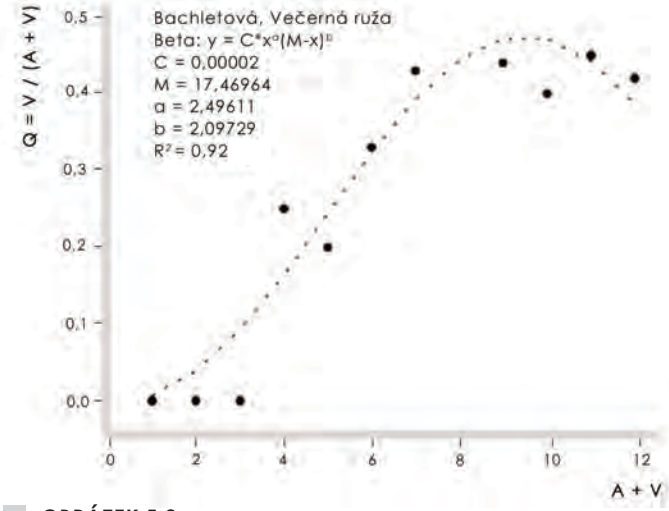
OBRÁZEK 5.2b
Průběh aktivity textu modelovaný prostřednictvím Morseho funkce v básni E. Bachletové *Istota*.



OBRÁZEK 5.2c
Průběh aktivity textu modelovaný prostřednictvím Morseho funkce v básni E. Bachletové *Dielo Stvoritela*.



OBRÁZEK 5.2d
Průběh aktivity textu modelovaný prostřednictvím beta funkce v básni E. Bachletové *Večerné ticho*.



OBRÁZEK 5.2e
Průběh aktivity textu modelovaný prostřednictvím beta funkce v básni E. Bachletové *Večerná ruža*.

5.3 KLASIFIKACE TEXTŮ PODLE JEJICH CELKOVÉ AKTIVITY A DESKRIPTIVITY

Měření aktivity a deskriptivity umožňuje texty klasifikovat. Tato klasifikace je založena na porovnání celkové hodnoty aktivity *Q* (v Tabulce 5.1 jsou tyto hodnoty označeny tučně). Výhodou níže popsaného postupu je zejména to, že můžeme texty nejen klasifikovat, ale zároveň také testovat signifikantnost aktivity či deskriptivity. Pro testy použijeme dva způsoby navržené Altmannem (1978) a Altmannovou a Altmannem (2008):

a) pro texty, které obsahují malý (tj. pomocí vzorce (5.9) lehce počitatelný) počet verb *V* a adjektiv *A*, testujeme hypotézu o vysoké aktivitě textu prostřednictvím binomického kritéria

$$P(X \geq V) = \sum_{x=V}^{n} \binom{n}{x} 0{,}5^n \quad , \tag{5.9}$$

tj. počítáme pravděpodobnost, s níž *X* je rovno nebo větší než pozorovaná hodnota *V*. Analogicky testujeme hypotézu týkající se vysoké deskriptivity

$$P(X \leq V) = \sum_{x=0}^{V} \binom{n}{x} 0{,}5^n \ , \tag{5.10}$$

tj. počítáme pravděpodobnost, s níž *X* je rovno nebo menší než pozorovaná hodnota *V*.

Pokud jsou hodnoty vypočítané na základě vzorců (5.9) a (5.10) menší než zvolená hladina významnosti (např. $\alpha \leq 0{,}05$), tak mluvíme o textu vyjadřujícím signifikantně vysokou aktivitu v případě (5.9) a textu vyjadřujícím signifikantně vysokou deskriptivitu v případě (5.10).

b) pro texty, které obsahují velký počet verb (*V*) a adjektiv (*A*) (zpravidla se počítá s počty většími než 60), testujeme hypotézu o vysoké aktivitě textu prostřednictvím asymptotického testu, tj.

$$X^2 = \frac{(V - A)^2}{V + A} \ , \tag{5.11}$$

což je chí-kvadrát test s jedním stupněm volnosti, nebo identicky

$$u = (2Q - 1)\sqrt{V + A}\ , \tag{5.12}$$

což je normální test. Takže platí $u^2 = X^2$.

Pro ilustraci aplikujme výše uvedený postup při analýze básně *Rozdelená bytosť*, která obsahuje pět adjektiv a patnáct verb. Na základě vzorce (5.1) vypočítáme aktivitu textu

$$Q = \frac{15}{15 + 5} = 0{,}75\ .$$

Jedná se o signifikantně aktivní text, nebo nikoliv? Za použití vzorce (5.9) testujeme nulovou hypotézu, podle níž tato báseň není signifikantně aktivní:

$$P\,(X \geq V) = \sum_{x=15}^{20} \binom{20}{x} 0{,}5^{20} = 0{,}0207\ .$$

Protože 0,0207 < 0,05, odmítáme nulovou hypotézu (pro hladinu významnosti $\alpha \leq 0{,}05$) a můžeme konstatovat, že báseň je signifikantně aktivní. Pokud testujeme stejnou hypotézu prostřednictvím (5.11), dostáváme

$$X^2 = \frac{(15 - 5)^2}{15 + 5} = 5\ .$$

Jelikož je výsledná hodnota $X^2 = 5$ vyšší než tabulková hodnota pro jeden stupeň volnosti (3,84), ($P = 0{,}0253$), stejně jako v předchozím případě odmítáme nulovou hypotézu. Za použití normálního testu (5.12) získáváme

$$u = (2 \cdot 0{,}75 - 1)\sqrt{15 + 5} = 2{,}2361\ ,$$

což je vyšší než kritická hodnota 1,96 (pro hladinu významnosti $\alpha \leq 0{,}05$), takže opět zamítáme nulovou hypotézu (navíc je evidentní, že $X^2 = u^2$).

Na základě tohoto postupu – konkrétně prostřednictvím binomického kritéria (5.10) – jsme analyzovali všechny básně z Tab. 5.1, výsledky jsou prezentovány v Tab. 5.2. Tabulka obsahuje pouze signifikantně aktivní a deskriptivní básně, ostatní básně z Tab. 5.1 můžeme označit za neutrální.

TABULKA 5.2
Signifikantně aktivní a deskriptivní básně E. Bachletové.

Signifikantně aktivní básně	
Iba neha;	*Čakáme šťastie;*
Rozdelená bytosť;	*Stály smútok pre šesť písmen;*
Rozťatá prítomnosť;	*Malé modlitby;*
Malý ošiaľ;	*Vo večnosti slobodná.*
Signifikantně deskriptivní básně	
Naše dejiny;	*Večerné ticho;*
To všetko je dar;	*Precitnutie.*

5.4 KLASIFIKACE TEXTŮ PODLE PRŮBĚHU VÝVOJE JEJICH AKTIVITY A DESKRIPTIVITY

Také analýza průběhu aktivity umožňuje texty klasifikovat dle jejich vztahu k této vlastnosti. Je přitom otázkou, kterou je možné vyřešit jen prostřednictvím dalšího výzkumu, zda míra aktivity a její průběh nějakým způsobem souvisí s růzností žánrů, autorstvím, vývojem jazykové akvizice apod.

Nyní k samotné klasifikaci: za krajní případy takovéhoto třídění lze považovat texty, v nichž se na jedné straně vyskytují pouze verba a žádná adjektiva a vice versa. Označme tedy texty jako

(Ia) *extrémně aktivní*, pokud má Q v průběhu celého textu hodnotu $Q = 1$;

(Ib) *extrémně deskriptivní*, pokud má Q v průběhu celého textu hodnotu $Q = 0$.

Pro další skupinu textů platí, že průběh aktivity začíná v jedné z extrémních poloh, tj. $Q = 1$ nebo $Q = 0$, poté dochází k poklesu (v případě $Q = 1$) či nárůstu aktivity (v případě $Q = 0$), přičemž hodnota aktivity nedosáhne rovnováhy $Q = 0{,}5$. Označme texty jako

(IIa) *převážně aktivní*, pokud text začíná s hodnotou aktivity $Q = 1$ a nikdy nepřesáhne hodnoty $Q \leq 0{,}5$;

(IIb) *převážně deskriptivní*, pokud text začíná s hodnotou aktivity $Q = 0$ a nikdy nepřesáhne hodnoty $Q \geq 0{,}5$.

Poslední skupinu tvoří texty, které v průběhu vývoje aktivity dosáhnou rovnovážného stavu $Q = 0{,}5$ a poté se dále vyvíjejí jakýmkoliv způsobem (tj. hodnota Q dále roste, klesá či osciluje kolem rovnováhy). Pro jednoduchost rozdělme tento typ textů pouze do dvou skupin (samozřejmě je možné provést jemnější dělení, pokud se to ukáže badatelsky užitečné) a označme texty jako

(IIIa) *aktivně rovnovážné*, pokud text začíná s hodnotou aktivity $Q = 1$ a v průběhu vývoje dosáhne hodnoty $Q = 0{,}5$;

(IIIb) *deskriptivě rovnovážné*, pokud text začíná s hodnotou aktivity $Q = 0$ a v průběhu vývoje dosáhne hodnoty $Q = 0{,}5$.

Tento typ klasifikace je z největší pravděpodobností použitelný pro relativně krátké texty (poetické texty jsou tedy dobrými kandidáty), protože u delších textů se dá předpokládat, že v průběhu vývoje aktivity dosáhnou hodnoty $Q = 0{,}5$. V případě poezie E. Bachletové dostáváme výsledky prezentované v Tab. 5.3.

TABULKA 5.3
Klasifikace poezie E. Bachletové podle aktivity textu.

Typ	Báseň
Ia. extrémně aktivní	*Zázrak*;
Ib. extrémně deskriptivní	*Naše dejiny*; *Každodennosť*; *To všetko je dar*;
IIa. převážně aktivní	*Rozťatá prítomnosť*; *Malý ošiaľ*; *Neopusť ma*; *Nepoznateľné*; *Hľadanie odpovedí*; *Keď dohorí deň*; *Ešte raz*; *Malé modlitby*; *Naše svetlo*; *Vo večnosti slobodná*;
IIb. převážně deskriptivní	*Z neba do neba*; *Nemám rada bielu*; *Ihly na nebi*; *Iba v modlitbe*; *Otázka*; *Náš chrám*; *Večerné ticho*; *Zbytočné srdce*;

IIIa. aktivně rovnovážné	*Iba neha*; *Tak málo úsmevu*; *Bez rozlúčky*; *Mladé oči*; *Čakáme šťastie*; *Dielo Stvoriteľa*; *Som iná*; *Čakanie na Boží jas*; *Do večnosti beží čas*; *Kým ich máme*; *Istota*; *Tiché verše*; *Iba život*; *Zasľúbenie jasu*; *Precitnutie*;
IIIb. deskriptivně rovnovážné	*Neha domova*; *Moje určenie*; *Spájanie*; *Dnešný luxus*; *Naše mamy*; *Vyznania*; *Rozdelená bytosť*; *Podobnosť bytia*; *Stály smútok pre šesť písmen*; *Nado mnou Ty sám*; *Návraty*; *Prvotný* sen; *Večerná ruža*; *Čas pre nádych vône*; *Smútok*;

6 Menzerathův zákon

Menzerathův zákon vyjadřuje vztah mezi délkou jazykových jednotek náležejících do různých, vedle sebe bezprostředně ležících jazykových rovin: čím delší je v jazyce nějaký konstrukt (např. slovo měřeno počtem slabik či morfémů), tím kratší jsou v průměru jeho konstituenty (v případě slova jde o slabiky či morfémy měřeny počtem fonémů) (srov. Altmann 1980; Cramer 2005; Hřebíček 1997; Menzerath 1928). Jedná se o jeden z nejvíce ověřených jazykových zákonů, přičemž jeho platnost byla testována také v genetice, sociologii, psychologii či muzikologii (srov. Altmann – Schwibbe 1989; Baixeries a kol.[i] 2012, Baixeries a kol.[ii] 2013; Boroda – Altmann 1991; Li 2012).

Menzerathův zákon je založen na předpokladu, že relativní změna délky konstituentu dy/y je úměrná změně délky konstruktu dx, takže platí

$$\frac{dy}{y} \approx dx \,. \tag{6.1}$$

Po dosazení koeficientu úměry v podobě jednoduché funkce $\left(b + \frac{c}{x}\right)$, kde b je jazyková konstanta a c vliv mluvčího, získáváme diferenciální rovnici

$$\frac{dy}{y} = -\left(b + \frac{c}{x}\right) dx \quad , \tag{6.2}$$

jejímž řešením je rovnice

$$y = ax^{-b}e^{-cx} \,, \tag{6.3}$$

kde x je délka konstruktu měřena v jeho konstituentech, y je průměrná délka konstituentu, a, b a c jsou parametry a e je Eulerova konstanta, báze přirozených logaritmů (její hodnota je přibližně 2,72). Zkrácená varianta tohoto vztahu, tj. $c = 0$, dává

$$y = ax^{-b} \,, \tag{6.4}$$

tj. potenční funkci, která se používá také v podobě Zipfova zákona a pro různé účely téměř ve všech vědách.

V následujících řádcích se nejdříve zaměříme na analýzu délky slov, které mají v naší analýze Menzerathova zákona roli konstituentu (kap. 6.1), následně budeme sledovat vlastnosti délky veršů, jež mají roli konstruktu (kap. 6.2). Nakonec věnujeme pozornost vztahu mezi oběma vlastnostmi vzhledem k Menzerathovu zákonu (kap. 6.3).

6.1 DÉLKA SLOVA

Problematika týkající se analýzy délky slova stojí v centru pozornosti jak lingvistů, tak matematiků již velmi dlouhou dobu (srov. Altmann 2013; Best 1998, 2001, 2006; Grzybek 2006; Köhler 2008; Mačutek – Wimmer 2013; Popescu a kol.[ii] 2013; Schmidt 1996; Smith 2012; Wang 2013; Zörnig 2013). Navzdory všemu úsilí rozsah problémů asociovaných s problematikou délky slova stále roste: na jedné straně kvantitativní analýzy dalších a dalších jazyků přinášejí nové poznatky o charakteru frekvenční distribuce slov obecně, na straně druhé se ukazuje, že mnoho nově definovaných vlastností jazyka – za všechny uveďme např. motivy (Köhler – Naumann 2008) či koncept tzv. plné valence (Čech – Pajas – Mačutek 2010, Čech 2013) – úzce souvisí právě s délkou slova. Připomeňme, že délka slova hrála ústřední roli v analýzách G. Zipfa (1935, 1949) a že se délka jazykových jednotek stala jednou z centrálních jazykových vlastností, s níž pracuje synergetická teorie jazyka (Köhler 1986, 2005b).

V naší analýze vycházíme z předpokladu, že (1) délka slova je fenoménem, jehož distribuce v textu není chaotická (tj. řídí se určitými zákony), a že (2) v rámci jednoho jazyka se délka slova (např. její distribuce, průměr atd.) liší v závislosti na typu textu, žánru, stylu, autorství atd. (Best 2001, 2006; Grzybek 2006; Popescu a kol.[i] 2009). Proto očekáváme, že u žánrově jasně vymezených textů jednoho autora, v našem případě jde o lyrickou poezii E. Bachletové, bude distribuce délky slova v jednotlivých textech vykazovat velmi podobné vlastnosti, což konkrétně znamená, že by měla vykazovat vlastnosti stejného modelu.

Pro testování tohoto předpokladu jsme zvolili metodu navrženou J. K. Ordem (1972), známou jako Ordovo kritérium.

6.1.1 Ordovo kritérium

Aplikace Ordova kritéria v rámci textové analýzy umožňuje charakterizovat texty graficky tak, že rozdíly mezi nimi je možné pozorovat opticky. Pro tento účel jsou navrženy dvě veličiny

$$I = \frac{m_2}{m'_1} \tag{6.5}$$

$$S = \frac{m_3}{m_2} \ , \tag{6.6}$$

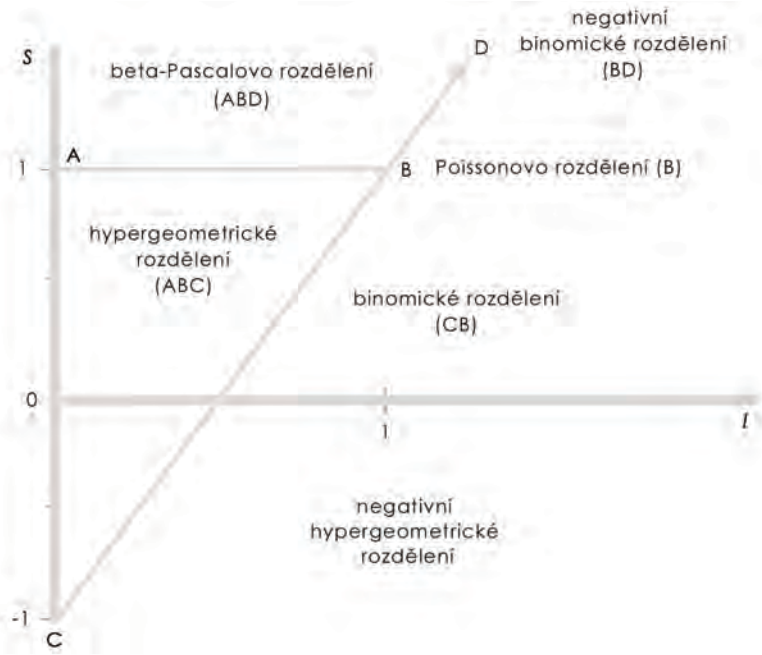
kde m'_1 je průměrná hodnota proměnné (v našem případě jde o délku slova), m_2 je druhý centrální moment (rozptyl) a m_3 třetí centrální moment (indikátor šikmosti distribuce) proměnné:

$$m'_1 = \frac{1}{N} \sum_{x=1}^{K} x f_x \ , \tag{6.7}$$

$$m_z = \frac{1}{N} \sum_{x=1}^{K} (x - m'_1)^z f_x \ , \tag{6.8}$$

kde x je délka slova, f_x je počet slov o délce x, N je součet všech frekvencí, z je řád daného centrálního momentu (v našem případě tedy $z \in \{2, 3\}$) a K je počet tříd, tj. K se rovná počtu různých délek slov v textu. Veličiny I a S je tak možno chápat jako funkci vyjadřující určité vlastnosti frekvenční distribuce (v našem případě délky slov).

Pokud empiricky zjištěné hodnoty I a S zakreslíme do grafu, můžeme výsledky porovnat s grafickým znázorněním, které charakterizuje jednotlivá rozdělení, viz Obr. 6.1.



OBRÁZEK 6.1
Grafické znázornění jednotlivých rozdělení navržené Ordem (1972).

6.1.2 Distribuce délky slova

Délka slova v analyzovaných básních byla měřena počtem slabik. V každé jednotlivé básni byl spočítán počet slov s délkou (tj. počtem slabik) $x = 1, 2, 3, 4\ldots$ Slova, jejichž délka se rovná nule, tj. neslabičné předložky, nebyla započítávána, protože tvoří přízvučné celky se slovy následujícími, jde o tzv. proklitika.

Konkrétně v básni *Aby spriesvitnela* je 14 slov jednoslabičných, 26 slov dvouslabičných, 15 slov tříslabičných, 5 slov čtyřslabičných a 4 slova pětislabičná (srov. druhý sloupec Tab. 6.1, kde jsou uvedena tzv. frekvenční spektra básní, tj. počty slov o dané délce). Na základě vzorce (6.7) získáváme

$$m_1' = \frac{(1 \cdot 14) + (2 \cdot 26) + (3 \cdot 15) + (4 \cdot 5) + (5 \cdot 4)}{64} = 2{,}359375\ .$$

Ze vzorce 6.8 vypočítáme hodnoty

$$m_2 = \frac{[(1 - m'_1)^2 14] + [(2 - m'_1)^2 26] + [(3 - m'_1)^2 15] + \cdots + [(5 - m'_1)^2 4]}{64} =$$
$$= 1{,}198975$$

a

$$m_3 = \frac{[(1 - m'_1)^3 14] + [(2 - m'_1)^3 26] + [(3 - m'_1)^3 15] + \cdots + [(5 - m'_1)^3 4]}{64} =$$
$$= 0{,}989067 \ .$$

Následně

$$I = \frac{m_2}{m'_1} = \frac{1{,}198975}{2{,}359375} = 0{,}5082$$

a

$$S = \frac{m_3}{m_2} = \frac{0{,}989067}{1{,}198975} = 0{,}8249 \ .$$

Analogicky byly vypočítány hodnoty veličin *I* a *S* u 47 lyrických básní E. Bachletové, viz Tab. 6.1.

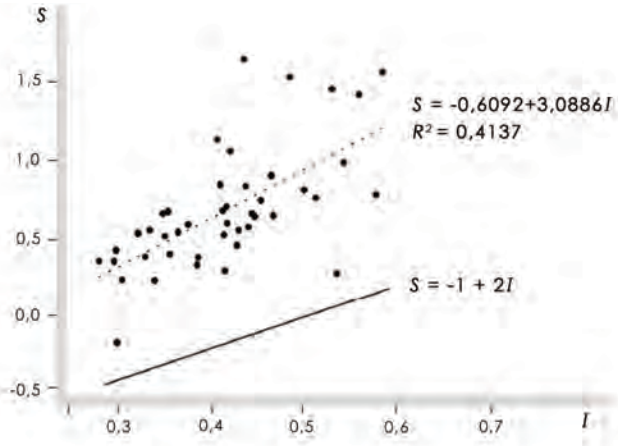
TABULKA 6.1
Hodnoty veličin *I* a *S* u 47 lyrických básní E. Bachletové. Hodnoty ve sloupci „distribuce" jsou seřazeny podle délky slova, tj. první hodnota vyjadřuje počet slov o délce 1, druhá hodnota počet slov o délce 2 atd., tj. jedná se o tzv. frekvenční spektrum.

Název	**Distribuce**	***I***	***S***
Aby spriesvitnela	14, 26, 15, 5, 4	0,5082	0,8249
Bez rozlúčky	15, 13, 5	0,3030	0,3739
Čakáme šťastie	8, 19, 9, 6, 2	0,4747	0,6625
Čakanie na Boží jas	30, 24, 18, 2	0,3938	0,3986
Čas pre nádych vône	23, 36, 16, 4, 0, 1	0,4284	1,0712

Název	**Distribuce**	***I***	***S***
Dielo Stvoriteľa	41, 53, 23, 13, 1	0,4545	0,6582
Dnešný luxus	14, 9, 8, 3, 1	0,5855	0,7951
Do večnosti beží čas	15, 22, 11, 1	0,3116	0,2563
Hľadanie odpovedí	22, 20, 18, 4	0,4223	0,3145
Iba neha	51, 53, 16, 5, 0, 2	0,4925	1,5436
Iba život	10, 18, 10, 5	0,3924	0,3497
Idem za Tebou	26, 28, 10, 5	0,4203	0,6943
Ihly na nebi	26, 18, 7, 1	0,3614	0,6887
Keď dohorí deň	22, 16, 11, 2	0,4215	0,5399
Kým ich máme	16, 20, 5, 1, 1	0,4145	1,1458
Len áno	7, 17, 5, 1	0,2867	0,3750
Malé modlitby	13, 25, 8, 2	0,3050	0,4421
Malý ošiaľ	32, 22, 12, 1	0,3721	0,5581
Mladé oči	7, 8, 3, 1	0,3830	0,6075
Moje určenie	55, 54, 25, 6, 2	0,4448	0,8493
Neopusť ma	6, 17, 7, 1, 0, 1	0,4432	1,6529
Náš chrám	28, 26, 19, 6, 1, 1	0,5509	0,9998
Naše dejiny	6, 7, 6, 5, 1	0,5435	0,2941
Naše mamy	22, 19, 11, 4	0,4481	0,5909
Naše svetlo	16, 23, 11, 7	0,4356	0,4740
Neha domova	10, 11, 3, 1	0,3556	0,6750
Nepoznateľné	39, 33, 12, 4, 1, 1	0,5381	1,4643
Podobnosť bytia	23, 32, 22, 5, 1, 1	0,4726	0,9170
Prvotný sen	23, 30, 13, 9, 2	0,5206	0,7757
Rozdelená bytosť	26, 31, 16, 3	0,3634	0,4184
Rozťatá prítomnosť	30, 34, 9, 3	0,3507	0,6667
Som iná	20, 24, 6, 4, 1, 1	0,5928	1,5737
Spájania	18, 14, 8, 2	0,4249	0,6133
Stály smútok pre šesť písmen	54, 64, 19, 4	0,3287	0,5505
Tak málo úsmevu	19, 25, 11, 5, 1	0,4614	0,7605
Tiché verše	8, 10, 11	0,3064	-0,1515

Název	Distribuce	*I*	*S*
To všetko je dar	16, 18, 12, 1	0,3469	0,2531
Večerná ruža	13, 18, 9, 2, 1, 1	0,5672	1,4309
Večerné ticho	25, 27, 9, 5	0,4242	0,7226
Vo večnosti slobodná	38, 71, 44, 5, 0, 1	0,3417	0,5724
Vrátili sa	17, 18, 9, 4	0,4375	0,5714
Vyznania	17, 25, 9, 3	0,3578	0,5331
Z neba do neba	13, 27, 18, 5, 0, 1	0,4174	0,8585
Zasľúbenie jasu	12, 23, 10, 3	0,3367	0,4026
Zbytočné srdce	12, 15, 5, 4	0,4517	0,6749

Pokud zaneseme výsledky prezentované v Tab. 6.1 do grafu kartézské soustavy souřadnic, získáme Obr. 6.2. Je patrné, že zjištěné hodnoty lze modelovat přímkou, přičemž rozptyl hodnot je relativně velký. Celková tendence může být vyjádřena vztahem $S = -0{,}6092 + 3{,}0886\,I$. Determinační koeficient ovšem vykazuje nízkou hodnotu $R^2 = 0{,}41$; předpokládáme, že nízká hodnota R^2 je způsobena zejména tím, že se jedná o poezii psanou volným veršem, což autorovi dává mnohem větší možnost „uniknout" mechanismům řídícím distribuci délky slova než v textech neuměleckých.



OBRÁZEK 6.2
Hodnoty <*I*, *S*> reprezentující distribuci délky slov v 47 básních E. Bachletové. Přímka $S = 2I - 1$ reprezentuje hranici beta-binomického rozdělení.

Navzdory tomu, že se data nacházejí v oblasti hypergeometrického rozdělení ($1 > S > 2I–1$; $I < 1$), není možné přímo tvrdit, že distribuce délky slov v poezii E. Bachletové lze modelovat právě tímto rozdělením, protože v dané oblasti se mohou nacházet také některá jiná rozdělení. První problém spočívá v charakteru dat: pro adekvátní modelování dat pomocí hypergeometrického rozdělení potřebujeme minimálně pět dobře reprezentovaných tříd délek, abychom získali alespoň jeden stupeň volnosti, který je nutný pro případné testování. Vzhledem k povaze dat musíme zvolit modely s menším počtem parametrů. Vhodnými kandidáty se jeví být Poissonovo ($N \to \infty, M \to \infty, n \to \infty, nM/N \to a$) a binomické rozdělení ($N \to \infty, M \to \infty, M/N \to p$), které představují limitní případy rozdělení hypergeometrického (srov. Obr. 6.1). Konkrétně Poissonovo rozdělení podle vzorce

$$P_x = \frac{a^{x-1}e^{-a}}{(x-1)!} \quad , \quad x = 1,2,\ldots \tag{6.9}$$

a binomické podle vzorce

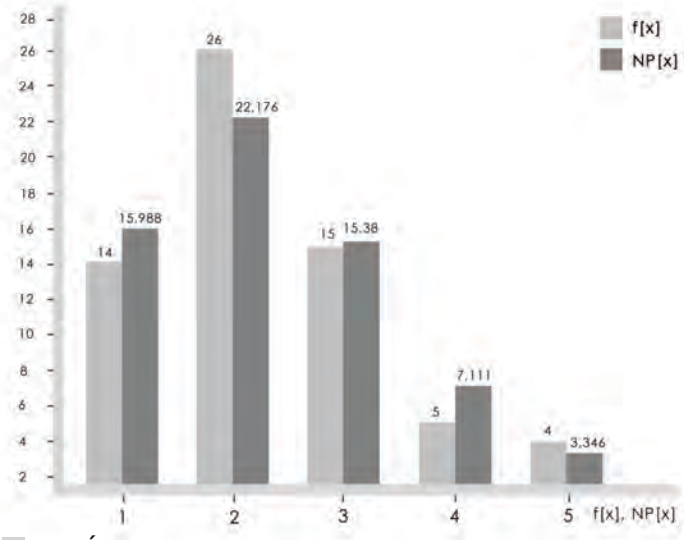
$$P_x = \binom{n}{x-1} p^{x-1} q^{n-x+1} \ , \ x = 1,2,\ldots n+1 \ . \tag{6.10}$$

V obou případech jsou rozdělení posunuta o jeden krok doprava, protože empirická data začínají vždy hodnotou $x = 1$. Parametry a a p se dají vypočítat buď iterativně, nejjednodušeji pomocí softwaru, např. *Altmann-Fitter – Iterative Anpasssung diskreter Wahrscheinlichkeitsverteilungen* (1994), nebo použitím standardních statistických odhadů. Parametr n zastupuje největší délku, ale je možné ho také vypočítat iterativně nebo odhadem. Jak je vidět například v básni *Aby spriesvitnela*, dostáváme pro binomické rozdělení úplně nereálnou hodnotu $n = 1374$ pro $p = 0{,}0010$. Narůstání n a zmenšování p je znakem toho, že binomické rozdělení konverguje k Poissonovu s parametrem $np = a$. V takovém případě můžeme binomické rozdělení jako model ignorovat. Jelikož v našich datech se tento jev vyskytuje v 18 případech (viz Tab. 6.2), můžeme Poissonovo rozdělení považovat za adekvátní model. Navíc skutečnost, že hypergeometrické rozdělení konverguje k Poissonovu, lze vnímat jako posilnění tohoto modelu.

Jelikož se dá parametr *a* Poissonova rozdělení odhadnout přímo z průměru, tj. $a = m_1'$, u posunutého rozdělení jako $m_1' - 1$, dostáváme v případě básně *Aby spriesvitnela* $a = 1{,}3594$. Iteračně je možné tento odhad zlepšit na $a = 1{,}3871$, který dává velice signifikantní shodu ($P = 0{,}64$) modelu s empirickými daty (srov. Tab. 6.2). S tímto parametrem získáváme teoretické hodnoty pro délky slov

$$NP_1 = 15{,}99$$
$$NP_2 = 22{,}18$$
$$NP_3 = 15{,}38$$
$$NP_4 = 7{,}11$$
$$NP_{\geq 5} = 3{,}35$$

Z důvodu testovaní shody pomocí chí-kvadrát testu jsme v tomto případě sloučili všechny třídy, jež jsou rovny nebo větší než $x \geq 5$ (výsledky tohoto testu jsou uvedeny v Tab. 6.2). Graficky je porovnání modelu s empirickými daty v případě básně *Aby spriesvitnela* znázorněno na Obr. 6.3.



OBRÁZEK 6.3
Porovnání modelu (Poissonovo rozdělení) pro distribuci délky slov s empirickými daty v případě básně *Aby spriesvitnela*.

V Tabulce 6.2 jsou uvedeny hodnoty chí-kvadrát testu dobré shody pro obě rozdělení. Výsledky jsou velmi přesvědčivé – všechny testy jsou bez výjimky signifikantní, tzn. že navržené modely dobře reprezentují pozorované distribuce. V některých případech bylo možné použít pouze Poissonovo rozdělení, protože počet tříd délek slov byl příliš malý pro modelování rozdělením binomickým (*Tiché verše, Bez rozlúčky*); v jednom případě bylo možné modelovat distribuci pouze binomickým rozdělením (*Vo večnosti slobodná*). Je třeba poznamenat, že v mnoha případech se binomické rozdělení blíží Poissonovu. To je evidentní tehdy, když je hodnota *n* velká a hodnota *p* velmi malá: pokud v těchto případech vynásobíme $n \cdot p$, získáváme hodnotu téměř identickou s hodnotou parametru *a* u Poissonova rozdělení.

TABULKA 6.2

Výsledky testování vhodnosti Poissonova a binomického rozdělení pomocí chí-kvadrát (X^2) testu dobré shody. Pro každou báseň jsou jednotlivá rozdělení uvedena v pořadí podle hodnoty *P*. Ve sloupci s označením *SV* je uveden počet stupňů volnosti.

Báseň	**Teoretické rozdělení**	**Parametry**	X^2	***SV***	***P***
Aby spriesvitela	Poissonovo binomické	a = 1,3871 n = 1374; p = 0,0010	1,67 1,67	3 2	0,64 0,43
Bez rozlučky	Poissonovo	a = 0,7408	0,26	1	0,61
Čakáme šťastie	binomické Poissonovo	n = 7; p = 0,3182 a = 2,2538	2,60 5,14	3 4	0,46 0,27
Čakanie na Boží jas	Poissonovo binomické	a = 0,9441 n = 6; p = 0,1529	4,48 3,70	2 1	0,11 0,05
Čas pre nádych vône	binomické Poison	n = 5; p = 0,2091 a = 1,0588	0,56 3,09	1 3	0,45 0,38
Dielo Stvoriteľa	Poissonovo binomické	a = 1,1161 n = 9; p = 0,1230	3,90 3,33	3 2	0,27 0,19
Dnešný luxus	Poissonovo binomické	a = 1,0914 n = 1088; p = 0,0010	1,81 1,82	2 1	0,40 0,18
Do večnosti beží čas	binomické Poissonovo	n = 3; p = 0,3216 a = 1,0206	0,30 4,00	1 2	0,59 0,14

Báseň	**Teoretické rozdělení**	**Parametry**	X^2	*SV*	*P*
Hľadanie odpovedí	Poissonovo binomické	a = 1,1189 n = 6; p = 0,1799	3,43 2,82	2 1	0,18 0,09
Iba neha	Poissonovo binomické	a = 0,8616 n = 864, p = 0,0010	2,17 2,16	3 2	0,54 0,34
Iba život	Poissonovo binomické	a = 1,2761 n = 4; p = 0,3130	0,95 0,58	2 1	0,62 0,44
Idem za Tebou	Poissonovo binomické	a = 0,9361 n = 925; p = 0,0010	0,63 0,62	2 1	0,73 0,43
Ihly na nebi	Poissonovo binomické	a = 0,6852 n = 7; p = 0,0969	0,40 0,24	2 1	0,82 0,63
Keď dohorí deň	Poissonovo binomické	a = 0,8926 n = 10; p = 0,0083	1,71 1,67	2 1	0,42 0,20
Kým ich máme	Poissonovo binomické	a = 0,8562 n = 6; p = 0,1421	1,99 1,22	2 1	0,37 0,27
Len áno	binomické Poissonovo	n = 3; p = 0,3337 a = 1,0219	1,84 5,56	1 2	0,18 0,06
Malé modlitby	binomické Poissonovo	n = 3; p = 0,3281 a = 1,0018	1,41 5,25	1 2	0,23 0,07
Malý ošiaľ	Poissonovo binomické	a = 0,7653 n = 5; p = 0,1493	2,27 2,01	2 1	0,32 0,16
Mladé oči	Poissonovo binomické	a = 0,9120 n = 904; p = 0,0010	0,26 0,26	2 1	0,88 0,61
Moje určenie	Poissonovo binomické	a = 0,9174 n = 13; p = 0,0707	0,41 0,28	3 2	0,94 0,87
Neopusť ma	binomické Poissonovo	n = 5; p = 0,2310 a = 1,1701	2,33 4,80	1 2	0,13 0,09
Náš chrám	Poissonovo binomické	a = 1,1173 n = 1123; p = 0,0010	0,91 0,91	3 2	0,82 0,63
Naše dějiny	Poissonovo binomické	a = 1,5794 n = 1548; p = 0,0010	1,53 1,52	3 2	0,68 0,47
Naše mamy	Poissonovo binomické	a = 0,9660 n = 958; p = 0,0010	0,26 0,26	2 1	0,88 0,61
Naše svetlo	Poissonovo binomické	a = 1,2015 n = 1181; p = 0,0010	0,51 0,51	2 1	0,77 0,48

Báseň	**Teoretické rozdělení**	**Parametry**	X^2	*SV*	*P*
Neha domova	Poissonovo binomické	a = 0,8167 n = 808; p = 0,0010	0,71 0,70	2 1	0,70 0,40
Nepoznateľné	Poissonovo binomické	a = 0,8493 n = 859; p = 0,0010	0,50 0,50	2 1	0,78 0,48
Podobnosť bytia	binomické Poissonovo	n = 7; p = 0,1701 a = 1,2003	0,81 2,05	2 3	0,67 0,56
Prvotný sen	Poissonovo binomické	a = 1,2049 n = 1194; p = 0,0010	1,93 1,93	3 2	0,59 0,38
Rozdelená bytosť	binomické Poissonovo	n= 4; p = 0,2375 a = 0,9810	0,14 2,31	1 2	0,70 0,31
Rozťatá prítomnosť	Poissonovo binomické	a = 0,8171 n = 4; p = 0,2040	2,55 1,31	2 1	0,28 0,25
Som iná	Poissonovo binomické	a = 0,8812 n = 1000; p = 0,0010	2,94 2,94	2 1	0,23 0,09
Spájania	Poissonovo binomické	a = 0,8757 n = 868; p = 0,0010	0,47 0,47	2 1	0,79 0,47
Stály smútok pre šesť písmen	binomické Poissonovo	n = 3; p = 0,2716 a = 0,8273	1,26 5,93	1 2	0,26 0,05
Tak málo úsmevu	Poissonovo binomické	a = 1,0914 n = 1089; p = 0,0010	0,79 0,79	3 2	0,85 0,67
Tiché verše	Poissonovo	a = 1,3057	0,01	1	0,92
To všetko je dar	binomické Poissonovo	n = 3; p = 0,3221 a = 1,0330	1,17 3,43	1 2	0,28 0,18
Večerná ruža	Poissonovo binomické	a = 1,0576 n = 9; p = 0,1234	0,57 0,11	2 1	0,75 0,74
Večerné ticho	Poissonovo binomické	a = 0,9329 n = 922; p = 0,0010	0,87 0,87	2 1	0,65 0,35
Vo večnosti slobodná	binomické	n = 5; p = 0,2287	6,12	2	0,05
Vrátili sa	Poissonovo binomické	a = 1,0244 n = 1015; p = 0,0010	0,01 0,01	2 1	0,99 0,92
Vyznania	binomické Poissonovo	n = 4; p = 0,2432 a = 0,9809	0,69 2,22	1 2	0,41 0,33

Báseň	Teoretické rozdělení	Parametry	X^2	*SV*	*P*
Z neba do neba	binomické Poissonovo	n = 5; p = 0,2555 a = 1,3094	0,58 4,10	2 3	0,78 0,25
Zasľúbenie jasu	binomické Poissonovo	n = 3; p = 0,3568 a = 1,0857	0,69 2,82	1 2	0,41 0,24
Zbytočné srdce	Poissonovo binomické	a = 1,0746 n = 1051; p = 0,0010	0,97 0,98	2 1	0,61 0,32

Shrnuto, distribuce délek slov se v lyrické poezii E. Bachletové řídí podle binomického rozdělení, přičemž Poissonovo rozdělení představuje jeho limitní hranici ($n \to \infty$, $p \to 0$, $np \to a$). Náš úvodní předpoklad – tj. že distribuce délky slova v jednotlivých textech stejného žánru by měla vykazovat velmi podobné vlastnosti, což konkrétně znamená, že by měla vykazovat vlastnosti stejného modelu – tedy nebyl vyvrácen. Připomínáme, že Popescu a kol.[i] (2009) ukázali, že hodnoty <*I*, *S*> se mohou výrazně lišit v závislosti na různých typech textu, žánru, autorství atd.

6.2 DÉLKA VERŠE

V některých typech poetických textů je délka verše určena konvencí daného typu poezie: např. je pevně určen počet slov, stop (např. hexametr), slabik (např. v sonetu) atd., které musí verš obsahovat. V těchto případech samozřejmě nemá smysl uvažovat o délce verše jako o proměnné. Podobné je to u některých jazyků s délkou slova: má-li jazyk jenom jednoslabičná slova, není ani délka slova proměnnou, která by se dala použít pro textovou analýzu. Jiná situace nastává u poezie psané veršem volným – zde se autor neřídí předem daným pravidlem, což mu umožňuje „svobodně" využívat délku verše jako dalšího uměleckého prostředku. Na druhou stranu se dá očekávat, že přes veškerou autorskou „svobodu" je také distribuce délky verše u tohoto typu poezie řízena nějakým mechanismem, který zohledňuje jak nároky uživatelů jazyka, tak i požadavky jazykového systému, jak je známe např. se synergetického modelu jazyka (srov. Köhler 1986, 2005). Tento mechanismus by se měl projevit pravidelnou distribucí délek verše.

Poznamenejme, že délka verše byla prostřednictvím metod kvantitativní lingvistiky analyzována jen zřídka, zřejmě jen Grotjahnem (1979) a Bestem (2012a, 2012b).

Stejně jako v případě distribuce délky slov budeme distribuci délek verše analyzovat u poezie E. Bachletové. Vzhledem k tomu, že některé básně jsou příliš krátké (tj. reprezentace jednotlivých frekvenčních tříd nemůže být spolehlivá), je nutné předem stanovit kritéria výběru. V našem případě jsme vybrali všechny básně, které mají alespoň 15 veršů a minimálně 4 frekvenční třídy (srov. Tab. 6.3).

Vyjdeme-li z Wimmerovy-Altmannovy obecné teorie (Wimmer – Altmann 2005), která bere do úvahy jak danost jazyka, tak i vliv mluvčího (autora) a posluchače (čtenáře), přičemž vztah mezi těmito faktory vyjadřuje pomocí diferenciální nebo diferenční rovnice, je zřejmé, že Poissonovo rozdělení vychází z diferenční rovnice

$$P_x = \frac{a}{x} P_{x-1} \ , \tag{6.11}$$

na základě čehož získáváme výše uvedený vzorec (6.9). Pro verše psané volným veršem je ovšem tento přístup nedostatečný, protože v něm není obsažen předpokládaný vliv posluchače (čtenáře), proto musí být přidán další modifikační parametr *b*:

$$P_x = \frac{a}{x + b - 1} P_{x-1} \ , \tag{6.12}$$

jejímž řešením získáváme

$$P_x = \frac{a^x}{b^{(x)}\,{}_1F_1(1; b; a)} \quad , \ x = 0,1,2,\dots \ , \tag{6.13}$$

kde je vzrůstající faktoriální funkce a ${}_1F_1(1; b; a)$ je konfluentní hypergeometrická funkce, která tvoří normovací konstantu. S ohledem na zřejmý fakt, že distribuce délek veršů neobsahuje $x = 0$, musí být vzorec (6.13) posunut o jeden krok vpravo, tj.

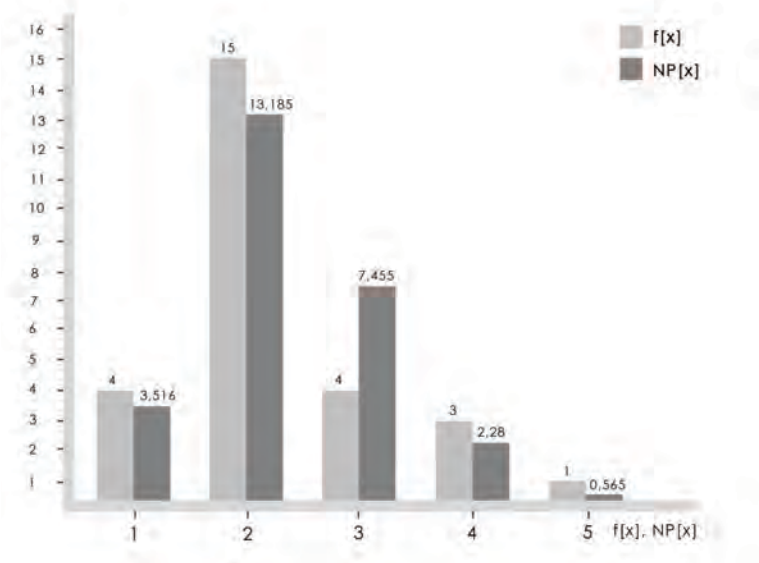
$$P_x = \frac{a^{x-1}}{b^{(x-1)}\,{}_1F_1(1; b; a)} \quad , \ x = 1,2,3,\dots \tag{6.14}$$

Takto získáváme posunuté hyper-Poissonovo rozdělení. Báseň *Čas pre nádych vône* neobsahuje ani jeden verš o délce jednoho slova, proto je v jejím případě rozdělení posunuto o dva kroky napravo.

Pro ilustraci uveďme srovnání modelu s pozorovanými hodnotami délek veršů v básni *Aby spriesvitnela*. Prostřednictvím softwaru *Altmann-Fitter* (1994) dostáváme parametry $a = 0{,}6658$ a $b = 0{,}1776$ a teoretické hodnoty pro jednotlivé délky veršů

$$NP_1 = 3{,}52$$
$$NP_2 = 13{,}19$$
$$NP_3 = 7{,}46$$
$$NP_4 = 2{,}28$$
$$NP_5 = 0{,}57$$

viz Obr. 6.4. Prostřednictvím chí-kvadrát testu zjišťujeme, že mezi modelem a pozorovanými délkami je významná shoda, protože $X^2 = 2{,}39$, $P = 0{,}12$, $SV = 1$.



OBRÁZEK 6.4
Porovnání modelu (hyper-Poissonovo rozdělení) pro délku veršů s empirickými daty v případě básně *Aby spriesvitnela*.

TABULKA 6.3

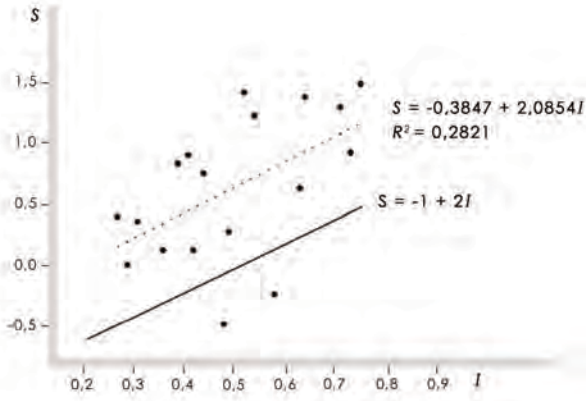
Délka veršů určena počtem slov v poezii E. Bachletové. V sloupci dva jsou uvedeny frekvence délky veršů v jednotlivých třídách, přičemž první hodnota odpovídá frekvenci veršů obsahující jedno slovo, druhá frekvenci veršů o dvou slovech atd. Parametry se vztahují k hyper-Poissonovu rozdělení. Ve sloupci s označením *SV* je uveden počet stupňů volnosti. Hodnoty *I* a *S* jsou veličiny pro aplikaci Ordova kritéria, viz kap. 6.1.2. a Obr. 6.5.

Báseň	**Frekvence délek v jednotlivých třídách**	**Parametr** ***a*** ***b***	X^2	***SV***	***P***	***I***	***S***
Aby spriesvitela	4, 15, 4, 3, 1	0,6658; 0,1776	2,39	1	0,12	0,41	0,92
Bez rozlúčky	4, 6, 5, 1	0,8294; 0,4813	0,73	1	0,39	0,36	0,15
Čakanie na Boží jas	7, 6, 10, 4, 1, 0, 1	1,9925; 1,5222	2,39	2	0,30	0,71	1,31
Čas pre nádych vône	0, 1, 3, 4, 2, 5, 2	2,6402; 0,8863	1,85	3	0,60	0,58	-0,21
Hľadanie odpovedí	5, 4, 6, 9	21,9759; 27,4687	1,64	1	0,20	0,48	-0,46
Iba neha	13, 15, 13, 9, 3, 1	1,9655; 1,4877	0,78	3	0,85	0,63	0,65
Ihly na nebi	2, 12, 3, 1, 3	0,7178; 0,1196	2,80	1	0,09	0,54	1,24
Malý ošiaľ	2, 14, 6, 5	0,7569; 0,1081	1,22	1	0,27	0,31	0,38
Moje určenie	10, 15, 12, 9, 2, 4	2,4553; 1,6677	1,93	3	0,59	0,73	0,94
Nepoznateľné	26, 16, 2, 6, 1	3,9217; 6,6769	6,96	2	0,03	0,64	1,39
Podobnosť bytia	4, 6, 10, 7, 1, 1	1,8377; 0,8675	2,65	3	0,45	0,49	0,30
Rozdelená bytosť	1, 9, 8, 5, 2, 1	1,2280; 0,1534	0,17	2	0,92	0,44	0,77
Som iná	1, 7, 10, 2, 1	0,9604; 0,1372	1,96	2	0,37	0,27	0,42

Báseň	**Frekvence délek v jednotlivých třídách**	**Parametr** *a* *b*	X^2	*SV*	*P*	*I*	*S*
Stály smútok pre šesť písmen	4, 13, 13, 15, 2, 1	1,6148; 0,4969	5,41	3	0,14	0,42	0,15
Tak málo úsmevu	1, 3, 10, 4, 2	1,6701; 0,5567	4,53	2	0,10	0,29	0,03
Vo večnosti slobodná	8, 22, 15, 6, 5, 2, 2	2,1271; 1,3296	4,20	3	0,24	0,75	1,50
Vyznania	7, 13, 4, 0, 2	0,5561; 0,2994	0,51	1	0,48	0,52	1,43
Z neba do neba	22, 12, 5, 1	0,8735; 1,5399	0,20	1	0,65	0,39	0,85

V případě některých básní je i Poissonovo rozdělení, jež je speciálním případem hyper-Poissonova rozdělení (když $b = 1$), akceptovatelným modelem (srov. Tab. 6.3). V jednom případě jsme nuceni použít limitní případ hyper-Poissonova rozdělení, konkrétně geometrické rozdělení, pro $a \to \infty$, $b \to \infty$, $\frac{a}{b} \Rightarrow q$, kde q je parametr geometrického rozdělení $P_x = pq^x, x = 0,1,2 \ldots$ V básni *Nepoznateľné* může být distribuce délek veršů modelována spíše geometrickým rozdělením s parametrem $p = 0{,}5517$ $(X^2 = 1{,}33, SV = 1, P = 0{,}27)$ nebo prostřednictvím Poissonova rozdělení s parametrem $a = 0{,}7134$, $(X^2 = 0{,}31, SV = 1, P = 0{,}57)$. Pro všechny ostatní básně je vhodný model hyper-Poissonova rozdělení. U básně *Hľadanie odpovedí* je zřejmý nárůst hodnot parametrů, což však neznamená nutnost změny modelu. Porovnáme-li distribuci délek veršů s délkou slov na základě Ordova kritéria (Obr. 6.5), vidíme, že hodnoty $\langle I,S \rangle$ délky veršů leží ve stejné oblasti jako hodnoty délky slov, i když rozptyl hodnot je evidentně větší než u délky slov. Lze očekávat, zejména vzhledem k velkému rozptylu, že v případě většího počtu textů by hodnoty $\langle I,S \rangle$ měly podobu elipsy.

Tyto výsledky lze chápat jako potvrzení našeho předpokladu, že distribuce délek veršů v poezii E. Bachletové není náhodná, ale že je projevem určitého mechanismu, pravděpodobně se jedná o projev autorského stylu.



OBRÁZEK 6.5
Hodnoty <*I*, *S*> reprezentující distribuci délky veršů na základě Tab. 6.3. Přímka *S* = 2*I* – 1 reprezentuje hranici beta-binomického rozdělení.

6.3 VZTAH DÉLKY VERŠE A DÉLKY SLOVA

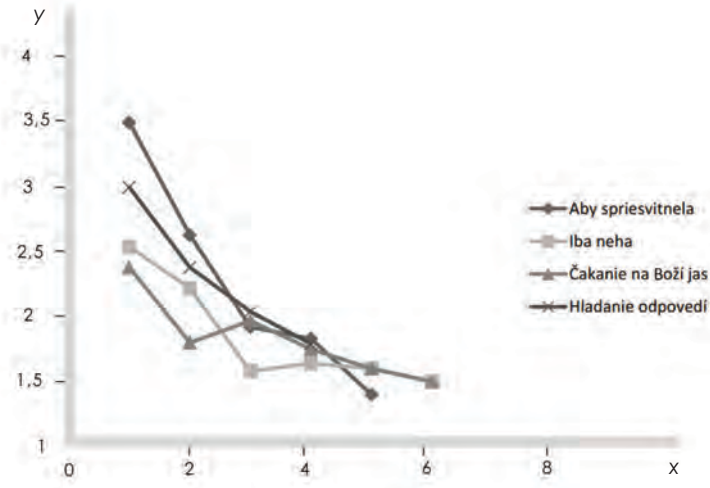
Specifický charakter básnických textů psaných volným veršem dovoluje vnímat volný verš jako textově-jazykovou jednotku, která má z hlediska fungování jazykového systému zvláštní postavení – na jedné straně je její hranice určena nejazykovým faktorem, tj. grafikou, na straně druhé je evidentní, že volba verše (a jeho délky) je do značné míry řízena ryze jazykovými faktory, ať už zvukovými, rytmickými, významovými atd. V tomto ohledu je tedy možné vnímat verš jako jednotku jazykovou a jeho délku jako jednu z jeho vlastností. V kap. 6.2 bylo ukázáno, že distribuce délek veršů, určena počtem slov, není náhodná a lze ji relativně dobře modelovat pomocí hyper-Poissonova rozdělení. Tyto výsledky nás opravňují předpokládat, že vztah verše a slova lze vnímat jako vztah konstruktu a konstituentu, tudíž můžeme očekávat, že vztah mezi délkou veršů a slov by měl být řízen Menzerathovým zákonem. Naše uvažování navíc vychází z jednoduché myšlenky: pokud básník použije jednoslovný verš, s největší pravděpodobností se bude jednat o slovo autosémantické (které je v naprosté většině případů delší než slovo synsémantické); s narůstající délkou verše roste pravděpodobnost, že jeho součástí budou také slova synsémantická, tudíž průměrná délka slov ve verši se bude zkracovat.

Stejně jako v případě analýzy distribuce délek verše není možné ani pro testování této hypotézy použít básně s malým počtem veršů. Proto jsme vybrali 14 básní, jejichž vlastnosti dovolují test provést, srov. Tab. 6.4. Ačkoliv u některých básní není počet případů v rámci jednotlivých tříd délek verše reprezentativní, výsledky ukazují na obecnou tendenci danou Menzerathovým zákonem, $y = ax^{-b}$; hodnoty parametrů *a*, *b* a determinačního koeficientu R^2, jsou uvedeny v posledních třech sloupcích Tab. 6.4. Graficky je vztah délky verše a délky slova na Obr. 6.6–6.9.

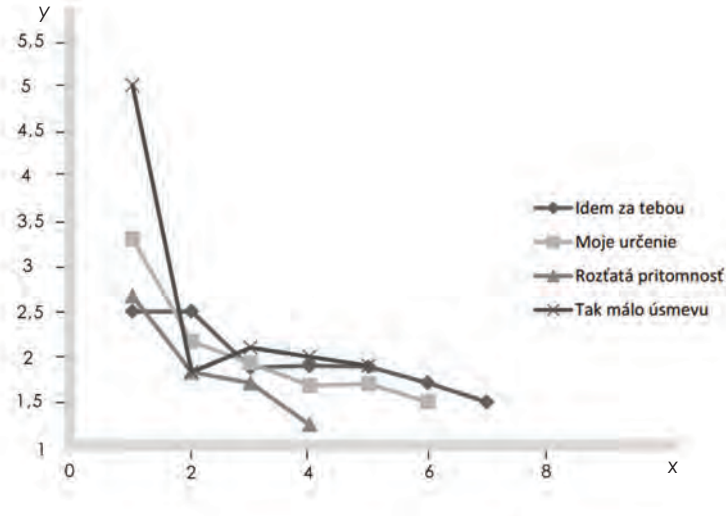
TABULKA 6.4

Vztah délky verše a délky slova v poezii E. Bachletové.

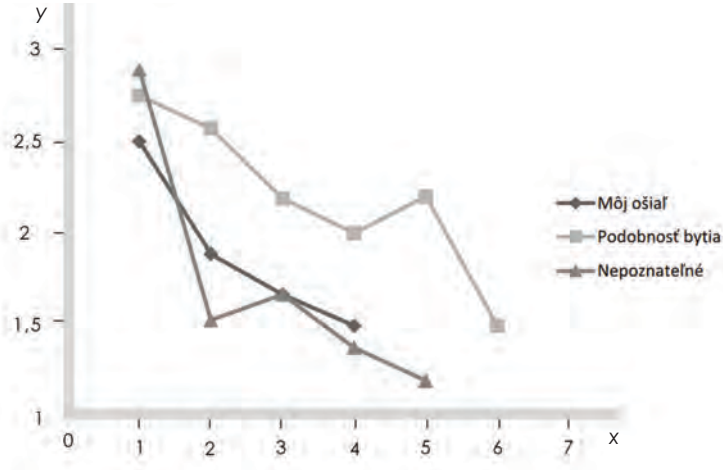
Báseň	**Délka verše (počet slov) *x***								**a**	**b**	**R^2**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**			
	Průměrná délka slova (počet slabik) *y*										
Aby spriesvitnela	3,5	2,63	1,92	1,83	1,4				3,5534	0,5215	0,98
Iba neha	2,54	2,22	1,58	1,64	1,6	1,5			2,5586	0,3147	0,91
Čakanie na Boží jas	2,38	1,8	1,97	1,75	1,6	1,5			2,3303	0,2281	0,86
Hladanie odpovedí	3	2,38	2,04	1,79					2,4823	0,0174	0,80
Idem za tebou	---	2,5	2,5	1,88	1,9	1,89	1,71	1,5	3,3009	0,3446	0,86
Moje určenie	3,3	2,17	1,93	1,68	1,7	1,5			3,2094	0,4466	0,97
Rozťatá prítomnosť	2,67	1,83	1,71	1,25					2,6658	0,4920	0,96
Tak málo úsmevu	5	1,83	2,1	2	1,9				4,6872	0,7409	0,81
Môj ošiaľ	2,5	1,89	1,67	1,5					2,4883	0,3703	~1,0
Podobnosť bytia	2,75	2,57	2,19	2	2,2	1,5			2,8539	0,2487	0,78
Nepoznateľné	2,89	1,53	1,67	1,38	1,2				2,7772	0,5479	0,90
Dielo Stvoriteľa	3,5	2,54	2,06	1,97	1,67	1,33			3,4948	0,4542	0,99
Z neba do neba	2,73	2,13	1,78						2,7380	0,3813	~1,0
Rozdelená bytosť	4	2,59	1,53	1,8	1,6	1,33			3,9520	0,6328	0,95



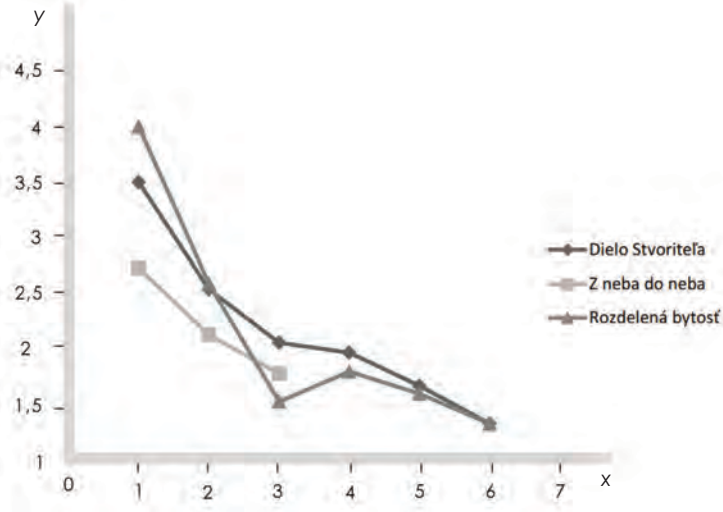
OBRÁZEK 6.6
Vztah délky verše (osa x) a průměrné délky slova (osa y) ve vybraných básních E. Bachletové (srov. Tab. 6.4).



OBRÁZEK 6.7
Vztah délky verše (osa x) a průměrné délky slova (osa y) ve vybraných básních E. Bachletové (srov. Tab. 6.4).



OBRÁZEK 6.8
Vztah délky verše (osa x) a průměrné délky slova (osa y) ve vybraných básních E. Bachletové (srov. Tab. 6.4).



OBRÁZEK 6.9
Vztah délky verše (osa x) a průměrné délky slova (osa y) ve vybraných básních E. Bachletové (srov. Tab. 6.4).

Zjištěné výsledky, na základě kterých nemůžeme odmítnout hypotézu o vztahu mezi délkou verše (jako konstruktu) a délkou slova (jako konstituentu) vyjádřenou Menzerathovým zákonem, bezpochyby rozšiřují náš pohled na fungování jazyka. Pokud totiž tento zákon vyjadřuje vztah mezi jednotkou „čistě" lingvistickou (nahlíženo z perspektivy tradičního lingvistického modelu) a jednotkou textovou (jejíž fungování je ovlivněno např. i estetickými a uměleckými hledisky nejrůznějšího druhu), vyvstává otázka, zda jsou podobně řízeny i další vztahy: např. délka strof, dvojice rýmujících se veršů apod. Naše výsledky jsou dalším příkladem toho, že vztah jazyka a textu je mnohem provázanější, než by se mohlo z tradičního podhledu na fungování a vztah obou entit zdát (srov. Hřebíček 1997, 2002).

7. Eufonie

Eufonie je jazykový prostředek, o jehož významu zejména v básnických textech nikdo nepochybuje. Tradičně bývá označována jako zvukový efekt, který vzniká záměrným uspořádáním sledu hlásek, přičemž „estetická působivost hlásek má svůj zdroj v seřadění, kterým se na ně upozorňuje" (Mukařovský 1940: 123). Sledujeme-li ovšem bližší charakterizaci tohoto jazykového jevu, vidíme, že jasná kritéria pro označení eufonického sledu hlásek obvykle chybějí. Pro ilustraci uveďme slova jednoho z předních českých literárních teoretiků 20. století J. Mukařovského (1940: 123):

(1) „Eufonické uspořádání sledu hlásek děje se nejčastěji tak, že se jistá hláska mnohonásobně opakuje nebo že se opakuje jednou, po případě i několikrát celé jisté seskupení hlásek v sestavě buď stejné nebo poněkud obměněné."

(2) „K *nahodilým* seskupením stejných hlásek nebo i celých opakujících se hláskových skupin docházívá pro omezenost hláskového repertoáru – tak na př. při pouhých pěti českých samohláskách – i v tekstech postrádajících eufonické, ba dokonce estetické záměrnosti, avšak takováto seskupení unikají zpravidla pozornosti čtenářově."

Podle jakých kritérií poznáme, kdy jde o eufonické uspořádání a kdy o uspořádání nahodilé? Kdy se jedná o estetickou záměrnost autora a kdy o nahodilost? A jak poznáme, která seskupení unikají pozornosti čtenáře a která nikoliv? Mukařovský poukazuje např. na kooperaci eufonie s rytmickými, syntaktickými a významovými charakteristikami textu, ale žádná jednoznačná kritéria neodhaluje; naopak, odkaz na vliv blíže nedefinovaných vlastností spíše celý přístup k eufonii činí ještě více jednoznačně neuchopitelným.

Vágní vymezení eufonie nacházíme i v moderních slovnících literární teorie, srov. definici eufonie v *Oxford Dictionary of Literary Terms* (Baldick 2008: 118) „A pleasing smoothness of sound, perceived by the ease with which the words can be spoken in combination. The use of long vowels, liquid consonants (*l, r*), and semi-vowels (*w, y*), contributes to euphony, along with the avoidance of adjacent stresses; the meaning of the words, however, has an important effect too. Euphony is the opposite of cacophony."

Pokud se chceme vyhnout subjektivnímu a velmi vágnímu hodnocení eufonie, máme zhruba tři možnosti, jak postupovat (srov. Wimmer – Altmann – Hřebíček – Ondrejovič – Wimmerová 2003: 56):

(1) zeptat se přímo autora, zda jev, který hodnotíme jako eufonii, takto skutečně zamýšlel použít, případně zda jej takto vnímá;
(2) zkoumat vnímání čtenářů;
(3) za použití statistických metod sledovat „nenáhodné" (tj. z hlediska pravděpodobnosti signifikantně vysoké) výskyty sledovaných zvukových prostředků.

Každá z výše uvedených metod má zřejmé nedostatky:

ad (1) zdá se rozumné předpokládat, že mnozí autoři používají eufonii spíše podvědomě, mnohdy jde o projev jazykového „citu" atd., takže autor sám si de facto není používání eufonie jako jazykového prostředku vůbec vědom, zejména když se soustředí primárně na obsahovou, nikoliv formální stránku textu; nelze také opomenout fakt, že mnoha autorů se nelze zeptat, protože již nežijí;
ad (2) shromáždit dostatečně velký a sociologicky vyvážený vzorek čtenářů není vůbec jednoduché. Navíc hodnocení založené na introspekci je vždy zatíženo vážnými problémy, srov. Meili (1967) či Estes (2000);
ad (3) statistické postupy se mnohým mohou jevit jako příliš redukcionistické, nedbající na komplexní povahu zkoumaných jevů.

Z výše uvedeného vyplývá, že zřejmě neexistuje žádná jediná „správná" metoda, jak eufonii zkoumat. V tomto duchu se přikláníme ke statistické metodě a máme pro to následující důvody:

a) intersubjektivita – výsledky analýz jsou nezávislé na osobě badatele. Jinak řečeno, pokud stejný test na stejném jazykovém materiálu provedou různí badatelé, dojdou k stejným výsledkům;

b) omezení vlivu náhody, byť data mnohdy působí intuitivně velmi přesvědčivě;

c) jasná formulace hypotézy týkající se eufonie může nejen potvrdit/vyvrátit, zda se daný jev v textu vyskytuje, ale umožňuje i jasně určit, co přesně eufonii v textu způsobuje.

7.1 METODA MĚŘENÍ EUFONIE

Vyjděme z předpokladu, že pro každou hlásku můžeme odvodit nějakou pravděpodobnost výskytu. Tato pravděpodobnost závisí na mnoha faktorech, zejména jde o vlastnosti jazyka samotného (např. počet hlásek, pravidla omezující možnosti jejich posloupnosti aj.), ale může jít i o faktory pragmatické (např. artikulační, žánrově podmíněné). Určit nějaké *obecně platné* pravděpodobnosti výskytu hlásek pro daný jazyk je však prakticky nemožné – je to dáno tím, že v jazyce není možné vymezit základní soubor, z něhož by se tyto obecně platné pravděpodobnosti odvodily. Proto je třeba k odvození pravděpodobnosti přistupovat s ohledem na konkrétní analýzu: někdy může sloužit pro odvození pravděpodobností jednotlivý žánr, autorský korpus, korpus jednoho časopisu atd. Vzhledem k tomu, že budeme analýzu eufonie ilustrovat na poezii E. Bachletové, použijeme jako základní soubor pro odvození pravděpodobnosti výskytu jednotlivých hlásek autorský korpus básnických textů napsaných touto autorkou.

Dále je třeba jednoznačně vymezit inventář hlásek. Pro analýzu poezie E. Bachletové jsou hlásky určeny následovně: dlouhé a krátké vokály jsou považovány za eufonicky identické, tj. např. hlásky [a] a [a:] jsou zapsány pouze jako [a]; vokál zapisovaný jako {ä} je ve shodě s výslovností autorky (osobní komunikace) zapisován jako [e]; jsou zapisovány čtyři diftongy [ia, ie, iu, uo]; sekvence písmen zapisována jako {ov, ou} je interpretována jako dvě samostatné hlásky [o] a [v]. Konsonanty zapisované jako {l, Ĺ} a {r, ŕ} jsou zapisovány jako [l] a [r]. Všechny zvukové varianty fonému *n* jsou sjednoceny pod jeden symbol [n]. Grafém {ch} je zapisován jako [x].

Pozorované relativní četnosti výskytů jednotlivých hlásek v poezii E. Bachletové jsou uvedeny v Tab. 7.1. Vzhledem k zásadně rozdílným zvukovým vlastnostem vokálů a konsonantů (a z toho plynoucích rozdílných funkcí v jazyce) budeme sledovat eufonii u obou typů těchto hlásek zvlášť. Konkrétně to znamená, že budeme pracovat s pozorovanými četnostmi spočítanými vždy v rámci každého typu hlásek, tj. vokálů a konsonantů.

TABULKA 7.1

Pozorované relativní četnosti výskytů jednotlivých hlásek v poezii E. Bachletové.

Hláska	Frekvence	Relativní frekvence	Relativní frekvence ve skupině konsonantů a vokálů	Hláska	Frekvence	Relativní frekvence	Relativní frekvence ve skupině konsonantů a vokálů
a	913	0,10722255	0,25396384	*n*	307	0,03605402	0,06239837
e	792	0,09301233	0,22030598	*s*	444	0,05214328	0,09024390
o	777	0,09125073	0,21613352	*z*	167	0,01961245	0,03394309
i	673	0,07903699	0,18720445	*l*	154	0,01808573	0,03130081
u	270	0,03170875	0,07510431	*r*	375	0,04403993	0,07621951
ia	38	0,00446271	0,01057024	*c*	193	0,02266588	0,03922764
ie	116	0,01362302	0,03226704	*ď*	74	0,00869055	0,01504065
iu	1	0,00011744	0,00027816	*š*	93	0,0109219	0,01890244
uo	15	0,0017616	0,00417246	*ž*	93	0,0109219	0,01890244
p	252	0,02959483	0,05121951	*č*	90	0,01056958	0,01829268
b	182	0,02137405	0,03699187	*dž*	8	0,00093952	0,00162602
f	73	0,00857311	0,0148374	*ľ*	141	0,01655901	0,02865854
v	283	0,03323547	0,05752033	*j*	197	0,02313564	0,04004065
w	54	0,00634175	0,01097561	*ň*	205	0,02407516	0,04166667
m	374	0,04392249	0,07601626	*k*	262	0,03076923	0,05325203
t	355	0,04169113	0,07215447	*g*	24	0,00281856	0,00487805
d	212	0,02489724	0,04308943	*h*	133	0,0156195	0,02703252
ts	85	0,00998238	0,01727642	*x*	83	0,0097475	0,01686992
dz	7	0,00082208	0,00142276				

V naší analýze eufonii definujeme jako funkci nenáhodného (tj. signifikantního) výskytu jedné nebo více hlásek ve verši, přičemž v daném verši se samozřejmě musí (ve shodě s tradiční definicí eufonie) konkrétní hláska vyskytnout minimálně dvakrát. Pro ilustraci postupu výpočtu pravděpodobnosti použijeme desátý verš básně *Aby spriesvitnela*, který je tvořen jediným slovem

počítača.

V tomto verši jsou 4 konsonanty a 4 vokály, přičemž konsonant [č] a vokál [a] se vyskytují ve verši dvakrát, jsou to tedy hlásky, jejichž prostřednictvím se v daném verši může projevit eufonie, jak jsme ji definovali výše. Zaměřme se nejdříve na konsonant [č] (u vokálu [a] bude postup analogický). Ptáme se, jaká je pravděpodobnost, že se na dvou místech, kde se může vyskytovat konsonant, vyskytne hláska [č]. Pokud označíme symbolem K jakýkoliv jiný konsonant než [č], mohou v daném verši nastat následující kombinace

ččKK, čKčK, čKKč, KččK, KčKč, KKčč.

Označíme-li pravděpodobnost výskytu [č] symbolem p, pravděpodobnost jiného konsonantu než [č] bude $q = 1 - p$, pak dostáváme následující možnosti

$ppqq, pqpq, pqqp, qppq, qpqp, qqpp$.

Pravděpodobnost výskytu [č] ve dvou pozicích ve verši pak vypočítáme

$$P([č] = 2) = ppqq + pqpq + pqqp + qppq + qpqp + qqpp = 6p^2q^2 .$$

Po dosazení hodnot pozorované relativní pravděpodobnosti hlásky [č] z Tab. 7.1 dostáváme

$$P([č] = 2) = 6p^2q^2 = 6\,(0{,}01829268)^2\,(1 - 0{,}01829268)^2 = 0{,}00193495 .$$

Situace však není tak jednoduchá, musíme totiž předpokládat, že se hláska [č] může vyskytnout jako prostředek eufonie nejen ve dvou, ale také ve třech i čtyřech pozicích ve verši. Pokud bychom analogicky k postupu pro dva výskyty vypsali všechny možnosti i pro tři a čtyři výskyty, dojdeme k zobecnění

$$(7.1) \qquad P([\check{c}] = x) = \binom{n}{x} p^x q^{n-x} \ ,$$

kde *x* je počet výskytů dané hlásky ve verši, *n* je celkový počet sledovaných prvků (tj. v našem případě konsonantů) ve verši a $\binom{n}{x}$ je binomickým koeficientem, který se dá vyjádřit jako

$$(7.2) \qquad \binom{n}{x} = \frac{n!}{x!\,(n-x)!} \ .$$

V našem případě tedy dostáváme

$$\begin{aligned} P([\check{c}] = 3) = \binom{4}{3} p^3 q^{4-3} &= \frac{4!}{3!(4-3)!}(0{,}01829268^3)(1 - 0{,}01829268) = \\ &= \frac{24}{6}(0{,}01829268^3)(1 - 0{,}01829268) = \\ &= 4(0{,}01829268^3)(1 - 0{,}01829268) = 0{,}00002404 \end{aligned}$$

a analogicky

$$P([\check{c}] = 4) = \binom{4}{4} p^4 q^{4-4} = 0{,}01829268^4 = 0{,}00000011 \ .$$

Pravděpodobnost, že se konsonant [č] ve verši vyskytne dvakrát až čtyřikrát, je dána součtem jednotlivých pravděpodobností, tj.

$$\begin{aligned} &P([\check{c}] = 2) + P([\check{c}] = 3) + P([\check{c}] = 4) = \\ &= 0{,}00193495 + 0{,}00002404 + 0{,}00000011 = 0{,}0019591 \ . \end{aligned}$$

Pokud za hladinu významnosti zvolíme α = 0,05, tak nastal jev, který má pravděpodobnost menší, než je zvolená hladina (0,0019591 < 0,05). Výskyt dvou konsonantů [č] v daném verši tedy považujeme za signifikantní. V souladu s naší definicí se tedy jedná o projev eufonie.

Celý postup můžeme zobecnit pomocí vzorce

$$(7.3) \qquad P(X \geq x_i) = \sum_{x=x_i}^{n} \binom{n}{x} p^x q^{n-x} \ ,$$

kde x_i je pozorovaná četnost hlásky ve verši (z hlediska naší definice eufonie je vždy $x_i > 1$).

V případě vokálu [a] tedy získáváme

$$P([a] \geq 2) = \sum_{x=2}^{4} \binom{4}{x} 0{,}25396384^{x} (1 - 0{,}25396384)^{4-x} = \ 0{,}2684252 \ ,$$

což je hodnota vyšší než zvolená hladina významnosti α = 0,05, tudíž se nejedná o signifikantní výskyt. Můžeme tedy konstatovat, že výskyty vokálu [a] v daném verši nejsou projevem eufonie.

Pro určení hodnoty eufonie hlásky ve verši je definován koeficient

$$E_{hláska} = \begin{cases} 100[\alpha - P(X \geq x_i)] & \text{pokud } \alpha > P(X \geq x_i) \\ 0 & \text{pokud } \alpha \leq P(X \geq x_i). \end{cases} \tag{7.4}$$

Takže například pro konsonant [č] má koeficient eufonie hodnotu

$$E_{[č]} = 100\,(0{,}05 - 0{,}0019591) = 4{,}8041 \ .$$

Eufonickou hodnotu celého verše definujeme jako průměr koeficientů těch hlásek, které se podle vzorce (7.4) nerovnají nule, tj.

$$E_{verš} = \frac{1}{k} \sum_{i=1}^{k} E_{hláska\ i} \ , \tag{7.5}$$

kde *k* je počet hlásek, u nichž se projevuje eufonie signifikantně (pozor, nejde o celkový počet hlásek ve verši). Následně můžeme definovat eufonickou hodnotu celé básně jako průměr eufonie veršů, tj.

$$E_{báseň} = \frac{1}{N} \sum_{j=1}^{N} E_{verš\ j} \ , \tag{7.6}$$

kde *N* je celkový počet veršů v básni. Pro všechny eufonické koeficienty pak platí, že leží v intervalu <0, 100α>, což jednoduše umožňuje porovnávat eufonické koeficienty jednotlivých hlásek, básní, sledovat průběh eufonie v textu atd.

Výpočet koeficientu eufonie celé básně ilustrujeme na příkladu básně *Aby spriesvitnela* (viz Tab. 7.2)

TABULKA 7.2

Hodnoty eufonie jednotlivých hlásek v básni *Aby spriesvitnela,* pro snadnější orientaci jsou v sloupcích 2 a 3 uvedeny celkové počty konsonantů (*K*) a vokálů (*V*) v daném verši.

Text	***K***	***V***	**Koeficent eufonie**
Nemám rada bielu	7	6	
dnes je prízrakom chladu	13	7	
znecitlivenia	7	5	[ň] = 1,8299
konečného verdiktu	10	7	
nad človekom	7	4	
nad pocitom	6	4	
nad láskou.	6	3	
Dnes je tu iná biela	8	6	
biela obrazovky	7	6	[b] = 2,4616
počítača	4	4	[č] = 4,8041
tam nahadzujeme	7	6	
svoje vnemy	6	4	
čiernymi linkami	8	6	[i] = 3,6665
rýchlo a bezpečne	8	6	
kreslíme životy	8	6	
slovami,	4	3	
ktoré navždy	7	4	
zmenili bielu	6	5	
a odviedli nás	6	5	
od základných farieb	10	6	
bytia.	2	2	
A možno stačí jedna	9	7	

Text	K	V	Koeficent eufonie
nenapísaná veta	7	7	
aby „novodobá“	5	6	[b] = 3,7301
biela spriesvitnela.	10	6	[l] = 1,2696, [ie] = 3,5678
Lebo čistá – biela krehkosť	13	8	
prichádza potichu…	7	6	[p] = 0,3620, [x] = 4,4351

Na základě vzorců (7.4) a (7.5) vypočítáme celkovou eufonii básně

$$E_{Aby\ spriesvitnela} = \frac{1}{27}\left[1{,}8299 + 2{,}4616 + 4{,}8041 + 3{,}6665 + 3{,}7301 + \right.$$
$$\left. + \frac{(1{,}2696 + 3{,}5678)}{2} + \frac{(0{,}3620 + 4{,}4351)}{2}\right] =$$
$$= \frac{21{,}3095}{27} = 0{,}7892\ .$$

Pro statistické testování rozdílů hodnot eufonie u jednotlivých básní je třeba znát rozptyl; ten vypočítáme podle vzorce

$$(7.7) \qquad Var(E_{báseň}) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\overline{E}_{verš\ i} - E_{báseň}\right)^2\ .$$

V případě *Aby spriesvitnela* dostáváme hodnotu

$$Var(E_{Aby\ spriesvitnela}) =$$
$$= \frac{1}{27-1}\left[(1{,}8299 - 0{,}7892)^2 + (2{,}4616 - 0{,}7892)^2 + (4{,}8041 - 0{,}7892)^2 + \right.$$
$$+ (3{,}6665 - 0{,}7892)^2 + (3{,}7301 - 0{,}7892)^2 + (2{,}4187 - 0{,}7892)^2 +$$
$$\left. + (2{,}3986 - 0{,}7892)^2 + 20(0 - 0{,}7892)^2\right] = 2{,}1011\ .$$

Pro porovnání hodnot eufonie celých básní použijeme asymptotický test

$$(7.8) \qquad |u| = \frac{E_{báseň\ 1} - E_{báseň\ 2}}{\sqrt{\frac{Var(E_{báseň\ 1})}{n_1} + \frac{Var(E_{báseň\ 2})}{n_2}}}\ ,$$

kde *n* je počet veršů v básni. Porovnáme-li např. eufonii básní *Aby spriesvitnela* a *Iba neha* (srov. Tab. 7.3), získáváme

$$|u| = \frac{0{,}7892 - 0{,}8698}{\sqrt{\frac{2{,}1011}{27} + \frac{2{,}5877}{54}}} = 0{,}2273 \ ,$$

což znamená, že mezi eufonií obou básní není signifikantní rozdíl.

TABULKA 7.3
Hodnoty eufonie v jednotlivých básních E. Bachletové,

Báseň	Počet veršů	Počet eufonických jevů	$E_{báseň}$	Var(E)
Aby spriesvitnela	27	9	0,7892	2,1011
Bez rozlúčky	16	6	0,7316	1,8172
Čakáme šťastie	13	9	1,1597	1,3579
Čakanie na Boží jas	29	5	0,4194	1,3195
Čas pre nádych vône	18	10	1,2001	2,6953
Dielo Stvoriteľa	44	19	0,7815	2,0019
Dnešný luxus	12	5	1,3363	3,4899
Do večnosti beží čas	18	5	0,5501	1,3292
Ešte raz	7	5	2,2522	3,2378
Hľadanie odpovedí	24	11	0,9947	2,1287
Iba neha	54	19	0,7784	2,2743
Iba v modlitbe	5	7	1,8158	1,3932
Iba život	14	29	2,6472	1,5591
Ihly na nebi	21	7	1,8567	1,0482
Istota	9	2	0,6610	1,9158
Každodennosť	8	6	1,9020	4,3369
Keď dohorí deň	14	10	1,6372	2,0889
Kým ich máme	16	3	0,4581	1,9220
Malé modlitby	11	21	2,4515	3,4998

Báseň	**Počet veršů**	**Počet eufonických jevů**	$E_{báseň}$	***Var(E)***
Mladé oči	7	2	0,6855	1,4395
Malý ošiaľ	27	12	0,8494	2,1483
Moje určenie	52	17	0,4907	1,1333
Nado mnou ty sám	10	4	0,9285	2,7892
Náš chrám	23	13	1,1775	2,9273
Naše dejiny	7	5	1,0391	2,0358
Naše mamy	14	4	0,9628	2,8418
Naše svetlo	28	17	1,5164	3,3362
Návraty	8	4	0,9654	2,7769
Neha domova	9	6	1,4756	4,0364
Neopusť ma	6	5	1,8638	2,3396
Nepoznateľné	51	19	0,6472	1,5851
Otázka	6	5	1,1409	2,1372
Podobnosť bytia	12	3	0,2917	0,3513
Precitnutie	13	6	0,8343	1,9732
Prvotný sen	27	15	1,1919	2,4199
Rozdelená bytosť	26	8	0,4548	1,0741
Rozťatá prítomnosť	36	8	0,5446	1,7323
Smútok	9	3	0,6482	1,2754
Som iná	21	5	0,4125	1,2397
Spájania	14	4	0,1713	0,2392
Stály smútok pre šesť písmen	48	15	0,4819	0,9211
Tak málo úsmevu	20	11	1,1373	2,5246
Tiché verše	12	2	0,3162	1,0939
To všetko je dar	24	8	0,6882	1,9045
Večerná ruža	15	10	1,4516	3,7202
Večerné ticho	19	39	2,6204	1,2604
Vo večnosti slobodná	31	9	0,8760	2,6253
Vrátili sa	12	4	0,5027	1,1785

Báseň	**Počet veršů**	**Počet eufonických jevů**	$E_{báseň}$	***Var(E)***
Vyznania	26	5	0,3935	1,3004
Z neba do neba	40	13	0,6763	1,7015
Zasľúbenie jasu	12	9	1,8127	2,5282
Zázrak	6	1	0,5695	1,9458
Zbytočné srdce	11	5	0,7625	1,3176

Hodnoty uvedené v Tab. 7.3 nevykazují žádné zřejmé pravidelnosti. Např. není možné postulovat vztah mezi délkou básně (měřenou v počtu veršů) a celkovou hodnotou eufonie. Předpokládáme, že tato nepravidelnost je do značné míry projevem typu poezie E. Bachletové, která je psána volným veršem. Hodnoty eufonie poezie E. Bachletové leží v intervalu <0,3162; 2,6472>, přičemž většina z nich je v dolní polovině intervalu <0; 5>, ve kterém se tyto hodnoty mohou nacházet. To je zřejmě důsledkem stylu autorky, která používá velmi krátké verše (mnohdy i jednoslovné), v nichž se větší eufonie ze zřejmých důvodů nemůže projevit. V případě rýmované poezie, kde zvuková stránka hraje vetší úlohu, se dá očekávat vyšší míra eufonie (toto tvrzení však musí být ověřeno experimentálně). Překvapivé je zjištění, že v analyzovaných textech není možné vysledovat ani vztah mezi jednotlivými hláskami a mírou eufonie. Dokonce ani vokály nemají na eufonii, jak jsme ji definovali, větší vliv než konsonanty, stejně tak není rozdíl mezi znělými a neznělými konsonanty.

Je třeba zdůraznit, že všechny zde uvedené závěry jsou jen prvním vhledem do kvantitativní analýzy eufonie a že je třeba je důkladně prozkoumat analýzou mnohem většího vzorku textů a autorů.

7.2 ALITERACE

Definujme aliteraci jako shodu hlásek na začátku veršů (nepůjde nám tedy o aliteraci v rámci jednoho verše). Tato stylistická figura je v poezii běžně používaná a stejně jako v případě eufonie není jasné, zda se jedná o prostředek vždy použitý vědomě, nebo zda jde o projev podvědomí (srov. Skinner 1939, 1941).

Pro kvantitativní analýzu aliterace je možné použít postup představený v kapitole 7.1, tj. budeme zkoumat, zda je výskyt hlásek na začátku veršů signifikantní, či nikoliv. Celou situaci ilustrujme opět na příkladu básně *Aby spriesvitnela*, ve které se v iniciálních pozicích veršů vyskytují následující hlásky: [ň, d, z, k, n, n, n, d, b, p, t, s, č, r, k, s, k, z, a, o, b, a, ň, a, b, ľ, p]. Z tohoto seznamu jednoduše spočítáme frekvence hlásek v iniciální pozici veršů

t	č	r	ľ	o	ň	d	z	p	s	k	n	b	a
1	1	1	1	1	2	2	2	2	2	3	3	3	3.

Pro výpočet aliterace použijeme vzorec (7.3), v němž *n* bude označovat počet veršů (v našem případě 27). Analogicky k postupu uvedenému v kap. 7.1 spočítáme hodnoty aliterace pro každou hlásku. V případě básně *Aby spriesvitnela* má pouze konsonant [b] signifikantně vysoké výskyty (A[b] = 3,0523), takže hodnota aliterace pro celou báseň je

$$A_{Aby\ spriesvitnela} = \frac{3{,}0523}{27} = 0{,}1131\ .$$

Hodnoty aliterace jednotlivých básní E. Bachletové jsou v Tab. 7.4.

TABULKA 7.4
Aliterace v poezii E. Bachletové,

Báseň	Počet veršů	Signifikantně aliterované hlásky	Hodnota aliterace celé básně
Aby spriesvitnela	27	b	0,1131
Bez rozlúčky	16	b, ž	0,2239
Čakáme šťastie	13	f, š	0,6617
Čakanie na Boží jas	29	a, n, p	0,3825
Čas pre nádych vône	18	-	0
Dielo Stvoriteľa	44	d, j, p, n	0,3261
Dnešný luxus	12	-	0
Do večnosti beží čas	18	u, l', b	0,591
Ešte raz	7	p	0,7024

Báseň	**Počet veršů**	**Signifikantně aliterované hlásky**	**Hodnota aliterace celé básně**
Hľadanie odpovedí	24	g	0,1996
Iba neha	54	a, t, č	0,1887
Iba v modlitbe	5	-	-
Iba život	14	z	0,1433
Ihly na nebi	21	n, ň	0,3829
Istota	9	u	0,2088
Každodennosť	8	u	0,3152
Keď dohorí oheň	13	p	0,339
Kým ich máme	16	h	0,1543
Malé modlitby	11	ň	0,4545
Malý ošiaľ	27	a	0,1078
Mladé oči	7	t	0,2609
Moje určenie	52	a, f, v, k	0,1766
Nado mnou Ty sám	10	p, d	0,4819
Náš chrám	23	a, p, v	0,2088
Naše dejiny	7	ď	0,6923
Naše mamy	14	-	-
Naše svetlo	28	j, k, d	0,4153
Návraty	8	d	0,4287
Neha domova	9	k	0,2276
Neopusť ma	6	ň	5,0000
Nepoznateľné	51	n, l'	0,1051
Otázka	6	-	0
Podobnosť bytia	29	z	0,1074
Precitnutie	13	b, h	0,4042
Prvotny sen	27	v, č, z,	0,4444
Rozdelená bytosť	26	ž, ň	0,167
Rozťatá prítomnosť	36	g, ň, p, ž	0,482
Smútok	9	-	0

Báseň	Počet veršů	Signifikantně aliterované hlásky	Hodnota aliterace celé básně
Som iná	21	s	0,2196
Spájanie	14	j	0,3306
Stály smútok pre šesť pismen	48	a, k, f, ň	0,3978
Tak málo úsmevu	20	s	0,2346
Tiché verše	12	b	0,1987
To všetko je dar	24	p, t, z	0,3774
Večerná ruža	15	j, p	0,2996
Večerné ticho	19	f	0,2603
Vo večnosti slobodná	31	č, ň	0,1372
Vrátili sa	12	f	0,3785
Vyznanie	26	t, ž	0,3585
Z neba do neba	40	d, p, x	0,2393
Zasľúbenie jasu	12	-	0
Zázrak	6	k, s	0,8582
Zbytočné srdce	11	d	0,1876

Kromě básně *Neopusť ma*, kde je aliterace absolutní, protože všechny verše začínají stejnou hláskou, nehraje aliterace v poezii E. Bachletové významnou roli, jak je vidět z Tab. 7.4.

Na závěr dodejme, že jak eufonie, tak i aliterace mohou být považovány za jeden z významných rysů autorského stylu a mohou být použity s jinými indexy např. v shlukové analýze zaměřené na identifikaci autorství.

Název softwaru

Altmann-Fitter – Iterative Anpasssung diskreter Wahrscheinlichkeitsverteilungen. Lüdenscheid: RAM-Verlag (1994).

Literatura

Texty **Evy Bachletové**, které byly použity pro analýzu, jsou dostupné na http://www.evabachletova.sk

Altmann, Gabriel

1966 „The measurement of euphony". In: *Teorie verše 1, Sborník brněnské versologické konference, 13.–16. května 1964*. Brno: Universita J. E. Purkyně, s. 208–209.

1978 „Zur Anwendung der Quotiente in der Textanalyse". *Glottometrika* 1, s. 91–106.

1980 „Prolegomena to Menzerath's Law". *Glottometrika* 2; s. 1–10.

1988 *Wiederholungen in Texten*. Bochum: Brockmeyer.

2006 „Fundamentals of quantitative linguistics". In: Genzor J. – Bucková, M. (eds.) *Favete linguis. Studies in Honour of Viktor Krupa*. Bratislava: Slovak Academic Press, s. 15–27.

2012 „Certain Differences between Qualitative and Quantitative Linguistics". *Czech and Slovak Linguistic Review* 2, s. 6–15.

2013 „Aspects of word length". In: Köhler, R. – Altmann G. (eds.) *Issues in Quantitative Linguistics* 3, Lüdenscheid: RAM, s. 23–38.

Altmann, Gabriel – Schwibbe, Michael H.

1989 *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Hildesheim: Georg Olms.

Altmann, Vivien – Altmann, Gabriel

2008 *Anleitung zu Quantitativen Textanalysen. Methoden und Anwendungen*. Lüdenscheid: RAM-Verlag.

Andres, Jan – Benešová, Martina

2011 „Fractal analysis of Poe's Raven". *Glottometrics* 21, s. 73–100.

Antosch, Friederike

1969 „The diagnosis of literary style with the verb-adjective ratio". In: Doležel, L. – Bailey, R. W. (eds.) *Statistics and style*. New York: Elsevier, s. 57–65.

Baayen, R. Harald

1989 *A corpus-based approach to morphological productivity. Statistical analysis and psycholinguistic interpretation*. Diss. Amsterdam: Free University.

Baixeries, Jaume – Hernández-Fernández, Antoni – Ferrer-i-Cancho, Ramon[i]

2012 „Random models of Menzerath–Altmann law in genomes". *BioSystems* 107 (3), s. 167–173.

Baixeries, Jaume – Hernández-Fernández, Antoni – Forns, Núria – Ferrer-i-Cancho, Ramon[ii]

2013 „The parameters of Menzerath-Altmann law in genomes". *Journal of Quantitative Linguistics* 20, s. 94–104.

Bakker, Franz J.

1965 „Untersuchungen zur Entwicklung des Aktionsquotienten". *Archiv für die gesamte Psychologie* 117, s. 78–101.

Baldick, Chris

2008 *Oxford Dictionary of Literary Terms*. Oxford University Press (3. vydání).

Bartoň, Tomáš – Cvrček, Václav – Čermák, František – Jelínek, Tomáš – Petkevič, Vladimír

2009 *Statistiky češtiny*. Nakladatelství Lidové noviny, Praha.

Benešová, Martina

2011 *Kvantitativní analýza textu se zvláštním zřetelem k analýze fraktální*. Olomouc (disertační práce). [Dostupné na: http://theses.cz/id/p19fdf/DISERTACNI_PRACE_BENESOVA.pdf]

Bernet, Charles

1988 „Faits lexicaux. Richesse du vocabulaire". In: Thoiron, P. et al. (eds.) *Etudes sur la richesse et la structure lexicale*. Paris: Champion, s. 1–11.

Best, Karl-Heinz

1998 „Results and perspectives of the Göttingen project on quantitative linguistics". *Journal of Quantitative Linguistics* 5, s. 155–162.

2006 „Sind Wort und Satzlänge brauchbare Kriterien zur Bestimmung von Lesbarkeit von Texten?". In: Wichter, S. – Busch, A. (eds.) *Wissenstransfer – Erfolgskontrolle und Rückmeldungen aus der Praxis*. Frankfurt am Main: Peter Lang Verlag, s. 21–31.

2012a „How many words are in a verse? An exploration". In: Naumann, S. – Grzybek, P. – Vulanovic, R. – Altmann, G., (eds.) *Synergetic linguistics. Text and language as dynamic systems*. Wien: Praesens, s. 13–22.

2012b „Zur Verslänge bei G. A. Bürger". *Glottometrics* 23, s. 57–62.

Best, Karl-Heinz (ed.)

2001 *Häufigkeitsverteilungen in Texten*. Göttingen: Peust & Gutschmidt Verlag.

Blatná, Renata

2001 „Básnický text a textový korpus". *Slovo a slovesnost* 62, s. 1–22.

Boder, David Pablo

1940 „The Adjective-Verb Quotient: A Contribution to the Psychology of Language". *Psychological Revue* 3, s. 309–345.

Boroda, Moisei G. – Altmann, Gabriel

1991 „Menzerath's law in musical texts". *Musikometrica* 3, s. 1–13.

Busemann, Adolf

1925 *Die Sprache der Jugend als Ausdruck der Entwicklungsrhythmik*. Jena: Fischer.

Bunge, Mario Augusto

1983 *Treatise on Basic Philosophy: Epistemology & Methodology*. Springer.

Bybee, Joan

2010 *Language, usage and cognition*. Cambridge: Cambridge University Press.

Comrie, Bernard

1989 *Language Universals and Linguistic Typology: Syntax and Morphology*. Chicago: University of Chicago Press (2. vydání).

Covington, Michael A. – McFall, Joe D.

2010 „Cutting the Gordian Knot: The Moving Average Type-Token Ratio (MATTR)". *Journal of Quantitative Linguistics* 17, s. 94–100.

Cramer, Irene M.

2005 „Das Menzerathsche Gesetz". In: Köhler, R. – Altmann, G. – Piotrowski, R. G. (eds.) *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook*, Berlin/New York: de Gruyter, s. 659–688.

Čech, Radek

2005a „Komunikace versus systém, nebo komunikace versus model?". *Slovo a slovesnost* 66, s. 176–179.

2005b „Limity (nejen jazykovědného) strukturalismu". *Slovo a slovesnost* 66 (1), s. 19–31.

2007 „Language System - Linguistics as an Empirical Science". *Sapostavitelno ezikoznanie* 32, s. 42–49.

2013 „Valence versus plná valence – obecnělingvistická analýza“. In: Gvoždiak, V. – Faltýnek, D. (eds.). *Tygrmatika*. Dokořán, s. 120–130.

2014a „Language and ideology: Quantitative thematic analysis of New Year speeches given by Czechoslovak and Czech presidents (1949–2011)“. *Quality & Quantity 48* (2), 899-910.

2014b „Text length and the lambda frequency structure of the text“. In: *Sequences in language and text* (přijato k publikaci).

Čech, Radek – Pajas, Petr – Mačutek, Ján

2010 „Full Valency. Verb Valency without Distinguishing Complements and Adjuncts“. *Journal of Quantitative Linguistics* 17, s. 291–302.

Čech, Radek – Altmann, Gabriel

2011 *Problems in Quantitative Linguistics 3*. Lüdenscheid: RAM-Verlag.

Čech, Radek – Popescu, Ioan-Iovitz – Altmann, Gabriel

2011a „Euphony in Slovak poetry“. *Glottometrics* 22, s. 5–16.

2011b „Word length in Slovak poetry“. *Glottometrics* 22, s. 44–56.

2013 „Methods of analysis of a thematic concentration of the text“. *Czech and Slovak Linguistic Review* (v tisku).

Čermák, František (ed.)

2007 *Frekvenční slovník mluvené češtiny*. Karolinum, Praha.

Čermák, František – Křen, Michal (eds.)

2004 *Frekvenční slovník češtiny*. Nakladatelství Lidové noviny, Praha.

Červenka, Miroslav – Sgallová, Květa

1997 „Verš a věta. Rytmické a větné členění v české poezii 2. pol. 19. století“. *Slovo a slovesnost* 58, s. 241–270.

Český národní korpus - SYN

2013 Ústav Českého národního korpusu FF UK, Praha. <http://www.korpus.cz>, datum přístupu: 1. 10. 2013.

Davidová Glogarová, Jana – Čech, Radek

2013 „Tematická koncentrace textu (kvantitativní analýza autorského stylu Ladislava Jehličky)". *Naše Řeč* 96, s. 234–254.

Davidová Glogarová, Jana – David, Jaroslav – Čech, Radek

2013 „Analýza tematické koncentrace textu – komparace publicistiky Ladislava Jehličky a Karla Čapka". *Slovo a slovesnost* 74, s. 41–54.

Doležel, Lubomír

1963 „Předběžný odhad entropie a redundance psané češtiny". *Slovo a slovesnost* 24, s. 165–174.

Doležel, Lubomír – Průcha, Jan

1966 „A statistical law of grapheme combinations". In: *Prague Studies in Mathematical Linguistics 1*, s. 33–43.

Ejiri, Koichi – Smith, Adolph E.

1993 „Proposal of a new 'constraint measure' for text". In: Köhler, R. – Rieger, B. B. (eds.) *Contributions to quantitative linguistics*. Dordrecht: Kluwer, s. 195–211.

Estes, William Kaye

2000 „Basic Methods in Psychological Science". In: Pawlik, K. – Rosenzweig, M. R. (eds.) *The International Handbook of Psychology*. London: SAGE Publications, s. 20–39.

Feyerabend, Paul Karl

2001 *Rozprava proti metodě*. Praha: Aurora.

Fraassen, Bas C. van

2002 *Empirical Stance*. New Haven & London: Yale University Press.

Fischer, Hardi

1969 „Entwicklung und Beurteilung des Stils". In: Kreutzer, H. – Gunzenhäuser, R. (eds.) *Mathematik und Dichtung*. München: Nymphenburger, s. 171–185.

Grotjahn, Rüdiger

1979 *Linguistische und statistische Methoden in Metrik und Textwissenschaft*. Bochum: Brockmeyer.

Grzybek, Peter (ed.)

2006 *Contributions to the science of text and language. Word length studies and related issues*. Dordrecht: Springer.

Guiraud, Pierre

1954 *Les catactères statistiques du vocabulaire*. Paris: Presses Universitaires de France.

1959 *Problèmes et mèthodes de la statistique linguistique*. Dordrecht: Reidel.

Hajič, Jan – Panevová, Jarmila – Hajičová, Eva – Pajas, Petr – Štěpánek, Jan – Havelka, Jiří – Mikulová, Marie

2006 *Prague Dependency Treebank 2.0*. Philadelphia: Linguistic Data Consortium.

Herdan, Gustav

1960 *Type-token mathematics*. The Hague: Mouton.

1966 *The advanced theory of language as choice and chance*. New York: Springer.

Hess, Carla W. – Sefton, Karen M. – Landry, Richard G.

1986 „Sample size and type-token ratios for oral language of preschool children". *Journal of Speech and Hearing Research* 29, s. 129–134.

Hess, Carla W. – Haug, Holy T. – Landry, Richard G.

1989 The reliability of type-token ratios for the oral language of school age children. *Journal of Speech and Hearing Research* 32, s. 536–540.

Hirsch, Jorge E.

2005 „An index to quantify an individual's research output". *Proceedings of the National Academy of Sciences of the USA* 102, s. 16569–16572.

Honoré, Antony

1979 „Some simple measures of richness of vocabulary". *ALLC Bulletin* 7, s. 172–177.

Hopper, Paul

1987 „Emergent Grammar". In: *Proceedings of the thirteenth annual meeting of the Berkley Linguistics Society*, 13, s. 139–157.

Hřebíček, Luděk

1997 *Lectures on Text Theory*. Praha: Oriental Institute.

2002 *Vyprávění o lingvistických experimentech s textem*. Praha: Academia.

Hudson, Richard

2007 *Language Networks: The New Word Grammar*. Oxford: Oxford University Press.

Jakobson, Roman

1995 „Gramatika poezie a poezie gramatiky". In: Červenka, M. (ed.) *Poetická funkce*. Praha: H&H.

Jelínek, Jaroslav – Bečka, Josef V. – Těšitelová, Marie

1961 *Frekvence slov, slovních druhů a tvarů v českém jazyce*. Praha: Státní pedagogické nakladatelství.

Kořenský, Jan

1987 „K procesuálnímu modelování řečové činnosti". *Slovo a slovesnost* 48, s. 177–189.

Köhler, Reinhard

1986 *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

2005 „Gegenstand und Arbeitsweise der Quantitativen Linguistik". In: Köhler, R. – Altmann, G. – Piotrowski, R. G. (eds.) *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook*, Berlin/New York: de Gruyter, s. 1–15.

2005b „Synergetic Linguistics". In: Köhler, R. – Altmann, G. – Piotrowski, R. G. (eds.) *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook*. Berlin – New York: de Gruyter, s. 760–774.

2008 „Word length in text. A study in the syntagmatic dimension". In: Mislovičová, S. (ed.) *Jazyk a jazykoveda v prohybe*. Bratislava: VEDA vydavatel'stvo SAV, s. 416–421.

2012 *Quantitative Syntax Analysis*. Berlin, New York: de Gruyter.

Köhler, Reinhard – Altmann, Gabriel

2005 „Aims and methods of quantitative linguistics". In: Altmann, G. – Levickij, V. – Perebyinis, V. (eds.) *Problemy kvantitativnoj lingvistiki*. Černivci: Ruta, s. 12–41.

2011 „Quantitative linguistics". In: Hogan, P. C. (ed) *The Cambridge Encyclopedia od the Language Sciences*, New York, Cambridge UP, s. 695–697.

Köhler, Reinhard – Naumann, Sven

2008 „Quantitative text analysis using L-, F- and T-segments". In: Preisach, B. – Schmidt-Thieme, D. (eds.) *Data Analysis, Machine Learning and Applications*. Berlin, Heidelberg: Springer, s. 637–646.

2010 „A syntagmatic approach to automatic text classification. Statistical properties of F- and L-motifs as text characteristics". In: Grzybek, P. – Kelih, E. – Mačutek, J. (eds.) *Text and Language*. Wien: Praesens Verlag, s. 81–89.

Krámský, Jiří

1942 „Příspěvek k fonologické statistice staré a nové angličtiny". *Časopis pro moderní filologii* 28, s. 376–384.

Krippendorff, Klaus

2012 *Content analysis: An introduction to its methodology*. Los Angeles – London – New Delhi – Singapore – Washington DC: SAGE Publications, Inc.. (third edition)

Lakoff, George

1973 „Fuzzy Grammar and the Performance/Comepetence Terminology Game". *Chicago Linguistic Society* 9, s. 271–291.

Levý, Jiří

1964 „Matematický a experimentální rozbor verše". *Česká literatura* 12, s. 181–212.

Li, Wentian

2012 „Menzerath's law at the gene-exon level in the human genome". *Complexity* 17 (4), s. 49–53.

Mačutek, Ján – Wimmer, Gejza

2013 „Alternative methods of goodness-of-fit evaluation applied to word length data". In: Köhler, R. – Altmann G. (eds.) *Issues in Quantitative Linguistics 3*, Lüdenscheid: RAM, s. 282–290.

Martinet, Andre

1964 *Elements of General Linguistics*. London: Faber and Faber Ltd.

Martynenko, Gregory

2010 „Measuring lexical richness and its harmony". In: Grzybek, P. – Kelih, E. – Mačutek, J. (eds.) *Text and Language. Structures · Functions · Interrelations. Quantitative Perspectives*. Wien: Praesens, s. 125–132.

McIntosh, Robert P.

1967 „An index of diversity and the relation of certain concepts to diversity". *Ecology* 48, s. 392–404.

Meili, Richard

1967 *Učebnice experimentální psychologie*. Praha: SPN.

Ménard, Nathan

1983 *Mesure de la richesse lexicale*. Paris: Slatkine.

Menzerath, Paul

1928 „Über einige phonetische Probleme". In: *Actes du premier Congres international de linguistes*. Leiden: Sijthoff, s. 104–105.

Miller, George A. – Beckwith, Richard – Fellbaum, Christiane – Gross, Derek – Miller, Katherine

1993 „Five Papers on WordNet". *CSL Report 43*, Cognitive Science Laboratory, Princeton University.

Mukařovský, Jan

1940 „O jazyce básnickém". *Slovo a slovesnost* 6 (3), s. 113–144.

Müller, Dieter

2002 „Computing the type token relation from the a priori distribution of types". *Journal of Quantitative Linguistics* 9, s. 193–214.

Neuendorf, Kimberly A.

2001 *The Content Analysis Guidebook*. Beverly Hills, CA: Sage Publications.

Ord, J. Keith

1972 *Families of Frequency Distributions*. London: Griffin.

Osgood, Charles E. – Walker, Evelyn G.

1959 „Motivation and language behavior: a content analysis of suicide notes". *Journal of abnormal and social psychology*, s. 58–67.

Panas, Espaminondas

2001 „The generalized Torquist: Specification and estimation of a new vocabulary text-size function". *Journal of Quantitative Linguistics* 8, s. 233–252.

Pieper, Ursula

1979 *Über die Aussagekraft statistischer Methoden für die linguistische Stilanalyse*. Tübingen: Narr.

Polanyi, Michael

1962 *Personal Knowledge. Towards a Post-Critical Philosophy*. Routlege: London.

Popescu, Ioan-Iovitz

2007 „Text ranking by the weight of highly frequent words". In: Grzybek, P. – Köhler, K. (eds) *Exact methods in the study of language and text (Quantitative Linguistics)*. Berlin – New York: Mouton de Gruyter, s. 557–567.

Popescu, Ioan-Iovitz – Altmann, Gabriel – Grzybek, Peter – Jayaram, Bijapur D. – Köhler, Reinhard – Krupa, Viktor – Mačutek, Ján – Pustet, Regina – Uhlířová, Ludmila – Vidya, Matummal N.[i]

2009 *Word frequency studies*. Berlin-New York: Mouton de Gruyter.

Popescu, Ioan-Iovitz – Mačutek, Ján – Altmann, Gabriel

2009 *Aspects of word frequencies*. Lüdenscheid: RAM.

2010 „Word forms, style and typology". *Glottotheory* 3/1, s. 89–96.

Popescu, Ioan-Iovitz – Altmann, Gabriel

2011 „Thematic concentration in texts". In: Kelih, E. – Levickij, V. V. – Matskulyak, Y. (eds.) *Issues in Quantitative Linguistics 2*, Lüdenscheid: RAM, s. 110–116.

Popescu, Ioan-Iovitz – Čech, Radek – Altmann, Gabriel

2010 „Structural conservatism and innovation in texts". *Glottotheory* 3/2, s. 43–64.

2011a „On stratification in poetry". *Glottometrics* 21, s. 54–59.

2011b „Vocabulary richness in Slovak poetry". *Glottometrics* 22, s. 62–72.

2011c *The lambda-structure of texts*. Lüdenscheid: RAM.

2012a „Some geometric properties of Slovak poetry". *Journal of Quantitative Linguistics* 19, s. 121–131.

2012b „Some characterizations of Slovak poetry". In: Naumann, S. – Grzybek, P. – Vulanović, R. – Altmann, G. (eds.) *Synergetic Linguistics. Text and Language as Dynamic Systems*. Wien: Praesens, s. 187–196.

2013 „Descriptivity in Slovak lyrics". *Glottotheory* 4, s. 92–104.

Popescu, Ioan-Iovitz – Naumann, Sven – Kelih, Emmerich – Rovenchak, Andrij – Overbeck, Anja – Sanada, Haruko – Smith, Reginald – Čech, Radek – Mohanty, Panchanan – Wilson, Andrew – Altmann, Gabriel[ii]

2013 „Word length: aspects and languages". In: Köhler, R. – Altmann G. (eds.) *Issues in Quantitative Linguistics 3*, Lüdenscheid: RAM, s. 224–281.

Ratkowsky, David A., Hantrais, Linda

1975 „Tables for comparing the richness and structure of vocabulary in texts of different length". *Computers and Humanities* 9, s. 69–75.

Schlissmann, Annemarie

1948 „Sprach und Stilanalyse mit einem vereinfachten Aktionsquotienten". *Wiener Zeitschrift für Philosophie, Psychologie und Pädagogik* 2, s. 41–62.

Schmidt, Peter (ed.)

1996 *Glottometrika 15: Issues in General Linguistic Theory and the Theory of Word Length*. Trier: WVT.

Schubert, Franziska

2008 *Differentielles Ausdrucksverhalten unter Berücksichtigung der Sprechersituation*. Dresden: Dissertationschrift.

Scott, Mike

2011 *WordSmith Tools version 6*, Liverpool: Lexical Analysis Software.

Skinner, Burrhus Frederic

1939 „The alliteration in Shakespeare's sonnets: A study of literary behaviour". *The Psychological Record* 3, s. 186–192.

1941 „A quantitative estimate of certain types of sound patterning in poetry". *The American Journal of Psychology* 54, s. 64–79.

Smith, Reginald

2012 „Distinct word length frequencies: distributions and symbol entropies". *Glottometrics* 23, s. 7–22.

Štukovský, Róbert – Altmann, Gabriel

1964 „Fonická povaha slovenského rýmu". *Litteraria* 7, s. 65–80.

1965 „Vývoj otvoreného rýmu v slovenskej poézii". *Litteraria* 8, s. 156–161.

Tešitelová, Marie

1968 „O básnickém jazyce z hlediska statistického". *Slovo a slovesnost* 29, s. 362–368.

1972 „On the so-called vocabulary richness". *Prague Studies in Mathematical Linguistics* 3, s. 103–120.

1992 *Quantitative Linguistics*. Prague: Academia; Amsterdam; Philadelphia: John Benjamins Publishing Co.

Těšitelová, Marie (ed.)

1985 *Kvantitativní charakteristiky současné spisovné češtiny*. Praha: Academia.

Thoiron, Philippe

1986 „Diversity index and entropy as measures of vocabulary richness". *Computers and the Humanities* 20, s. 197–202.

Trnka, Bohumil

1935 *A Phonological Analysis of Present-Day Standard English*. Praha: Universita Karlova.

1951 „Kvantitativní lingvistika". In: *Časopis pro moderní filologii* 34, s. 66–74.

Tuldava, Juhan

1995 „On the relation between text length and vocabulary size". In: Tuldava, J. (ed.) *Methods in quantitative linguistics*. Trier: WVT, s. 131–150.

2005 „Stylistics, author identification". In: Köhler, R. – Altmann, G. – Piotrowski, R. G. (eds.) *Quantitative Linguistics. An International Handbook*. Berlin-New York: de Gruyter, s. 368–387.

Tuzzi, Arjuna – Popescu, Ioan-Iovitz – Altmann, Gabriel

2010 *Quantitative analysis of Italian texts*. Lüdenscheid: RAM.

Tweedie, Fiona J. – Baayen, R. Harald

1998 „How variable may a constant be? Measure of lexical richness in perspective". *Computers and the Humanities* 32, s. 323–352.

Uhlířová, Ludmila

2005 „Quantitative linguistics in the Czech Republic". In: Köhler, R. – Altmann, G. – Piotrowski, R. G. (eds.) *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook*, Berlin/New York: de Gruyter, s. 129–135.

Wang Lu

2013 „Word length in Chinese". In: Köhler, R. – Altmann G. (eds.) *Issues in Quantitative Linguistics* 3, Lüdenscheid: RAM, s. 39–53.

Weitzman, Michael

1971 „How useful is the logarithmic type/token ratio?". *Journal of Linguistics* 7, s. 237–243.

Wilson, Andrew

2009 „Vocabulary richness and thematic concentration in internet fetish fantasies and literary short stories". *Glottotheory* 2 (2), s. 97–107.

Wimmer, Gejza – Altmann, Gabriel

2005 „Towards a unified derivation of some linguistic laws". In: Köhler, R. – Altmann, G. – Piotrowski, R.G. (eds.) *Quantitative Linguistics. An International Handbook*. Berlin-New York: de Gruyter, s. 307–316.

Wimmer, Gejza – Altmann, Gabriel – Hřebíček, Luděk – Ondrejovič, Slavomír – Wimmerová, Soňa

2003 *Úvod do analýzy textov*. Bratislava: Veda.

Yule, George Udny

1944 *The statistical study of literary vocabulary*. Cambridge: Cambridge University Press.

Ziegler, Arne – Altmann, Gabriel

2002 *Denotative Textanalyse*. Wien: Praesens.

Zipf, George Kingsley

1935 *The psycho-biology of language. An Introduction to Dynamic Philology*. Boston: Houghton-Mifflin. Cambridge: M.I.T. Press (2. vydání z roku 1968).

1949 *Human behavior and the principle of least effort*. Cambridge: Addison-Wesley. New York: Hafner (reprint z roku 1972).

Zörnig, Peter

2013 „Distances between words of equal length in a text". In: Köhler, R. – Altmann G. (eds.) *Issues in Quantitative Linguistics 3*, Lüdenscheid: RAM, s. 117–129.

Rejstřík věcný

V

Z

Rejstřík jmenný

G

H

J

K

L

M

Y

Z

Resumé

Metody kvantitativní analýzy (nejen) básnických textů

V knize jsou prezentovány metody, které lze použít pro kvantitativní analýzy textů; konkrétně jde o metody měření a) tematické koncentrace textu, b) slovního bohatství textu, c) aktivity textu, d) délky slova a jejího vztahu k délce verše (viz Menzerathův zákon) a e) eufonie. V rámci každé metody je také představen způsob, jak statisticky testovat rozdíly jednotlivými texty vzhledem k měřeným vlastnostem. Jednotlivé metodologické postupy jsou detailně popsány a tento popis je doplněn ilustrativním příkladem analýzy konkrétního textu. Kniha je uvedena kapitolou shrnující základní teoretická a metodologická východiska současné kvantitativní lingvistiky, a nabízí tak nejen prezentaci jednotlivých postupů, ale i jejich teoretického pozadí.

Quantitative text analysis metod (not only) for the poem

The book presents methods of quantitative text analysis; specifically, methods that can be used for measurement of a text's a) thematic concentration, b) vocabulary richness, c) activeness, d) word length and its relationship to the length of verse (see the Menzerath law), and e) euphony. For each individual method, a procedure is presented for statistical testing of differences between individual texts as per the qualities measured. These procedures are described in detail and accompanied each with illustrative examples of analyses of actual texts. The book opens with a chapter summarizing the basic theoretical and methodological foundations of contemporary quantitative linguistics and thus offers not only a presentation of the individual methods used therein, but also of their respective theoretical backgrounds.

Údaje o autorech

RADEK ČECH (1974)

Lingvista zabývající se kvantitativní textologií a kvantitativními analýzami syntaxe. Doposud se věnoval zejména analýzám tematické koncentrace a frekvenčních charakteristik textu. V oblasti syntaxe se zaměřuje na kvantitativní analýzy slovesné valence, tranzitivity a komplexních syntaktických sítí.

(více viz http://www.cechradek.cz)

IOAN-IOVITZ POPESCU (1932)

Jeden z nejvýznamnějších rumunských fyziků plazmatu, který se od roku 2006 zabývá kvantitativní lingvistikou. Navrhl množství indexů, pomocí nichž lze charakterizovat celou řadu vlastností textu. Za knihu „The Lambda Structure of Texts", kterou napsal ve spolupráci s G. Altmannem a R. Čechem, získal cenu Grigore C. Moisila v oblasti exaktních věd.

(více viz http://www.iipopescu.com)

GABRIEL ALTMANN (1931)

Původním zaměřením orientalista (vystudoval obory indonéština a japonština), který stál u zrodu kvantitativnělingvistického bádání, jehož hlavním cílem je překonat deskriptivní charakter analýz jazyka. Formuloval několik lingvistických zákonů, zabýval se obecnou teorií jazyka, aplikoval kvantitativnělingvistické metody při analýze všech jazykových rovin a textu. Od r. 2009 je čestným prezidentem Mezinárodní asociace kvantitativní lingvistiky (IQLA).

(více viz http://www.gabrielaltmann.de)

KATALOGIZACE V KNIZE - NÁRODNÍ KNIHOVNA ČR

Čech, Radek
Metody kvantitativní analýzy (nejen) básnických textů / Radek Čech, Ioan-Iovitz Popescu, Gabriel Altmann. -- 1. vyd. -- Olomouc : Univerzita Palackého v Olomouci, 2014. -- 136 s. -- (Qfwfq ; sv. 4)
České a anglické resumé

ISBN 978-80-244-4044-6 (brož.)

* 801.73 * 81'324
textová analýza
kvantitativní lingvistika
kvantitativní lingvistika -- metodologie
kolektivní monografie

81 - Lingvistika. Jazyky [11]

Metody kvantitativní analýzy (nejen) básnických textů
Radek Čech
Ioan-Iovitz Popescu
Gabriel Altmann

4. svazek Edice Qfwfq

Výkonný redaktor: Jiří Špička
Odpovědná redaktorka: Jana Kreiselová
Jazyková redakce: Martina Křížová
Sazba a obálka: Martina Šviráková

Vydala a vytiskla Univerzita Palackého v Olomouci
Křížkovského 8, 771 47 Olomouc
www.upol.cz/vup
e-mail: vup@upol.cz
Olomouc, 2014
1. vydání, 135 stran
čz 2014/220

ISBN 978-80-244-4044-6