Univerzita Palackého v Olomouci

Filozofická fakulta

Kvantitativní analýza textu

se zvláštním zřetelem k analýze fraktální

Martina Benešová

Disertační práce

Studijní program obecná lingvistika a teorie komunikace

Školitel: prof. RNDr. dr hab. Jan Andres, DSc.

Olomouc 2011

Prohlašuji, že jsem disertační práci zpracovala samostatně. Prohlašuji, že citace použitých pramenů je úplná a že jsem v práci neporušila autorská práva.

V Jihlavě, dne 31. 8. 2011 Martina Benešová

Děkuji

prof. RNDr. dr hab. Janu Andresovi, DSc. za cenné odborné a technické rady v průběhu sestavování práce a za podporu mého úsilí práci sepsat.

Děkuji

prof. PhDr. Janu Kořenskému, DrSc. za podporu při řešení tématu.

Martina Benešová

Obsah

Anotace

Kvantitativní analýza textu se zvláštním zřetelem k analýze fraktální

V této práci se snažím navázat na teze Ferdinanda de Saussura, Paula Menzeratha, Gabriela Altmanna a Luďka Hřebíčka. Zaměřuji se především na kvantitativní analýzu textu se speciálním důrazem na využití Menzerath-Altmannova zákona a teorie fraktálů. Tato analýza je na závěr doplněna kvantifikací dle teorie informace a numerické estetiky.

Jako analyzované výběrové soubory jsem zvolila jazykové a překladatelské mutace básně Edgara Allana Poea *The Raven*, tedy originální Poeův anglický text, osmnáct překladů do českého jazyka a jeden překlad do jazyka německého. Protože však jsou poetické texty značně specifické a náročné na kvantifikaci, pro kontrast jsem připojila též jako další výběrový soubor jeden žurnalistický text.

Práce je uvedena krátkým exkurzem do historie používání matematických metod v lingvistice a do historie matematické a kvantitativní lingvistiky. Důležitou součástí je vybudovaný algoritmus kvantitativního zpracování textu, který je doplněn praktickými aplikacemi, ukázkami výpočtů, tabulkami, obrázky a grafy. Je vybudována teorie jazykových fraktálů a stanoven způsob výpočtu stupně sémantičnosti textu. Modely prokázaných jazykových fraktálů jsou vizualizovány s pomocí teorie fraktálů společně s přidruženými matematickými fraktály. Vhodnost či nevhodnost kvantitativní analýzy poetických textů a volba jednotek pro analýzu je komparativně zkoumána a traktována.

Annotation

Quantitative Analysis of Text with Special Respect to Fractal Analysis

In my thesis, I try to follow up with the works of Ferdinand de Saussure, Paul Menzerath, Gabriel Altmann and Luděk Hřebíček. I focus especially on the fractal analysis of text with a special emphasis on the usage of the Menzerath-Altmann law and the theory of fractals. This analysis is, then, supported with the one using the information theory and numerical aesthetics.

I chose language and translation mutations of Edgar Allan Poe's poem *The Raven* to become samples for the analysis; i.e. they are the Poe's original English text, eighteen translations into the Czech language and one translation into the German language. Poetic texts are, nevertheless, considerably specific and demanding for quantifying, therefore I decided to add one more sample to contrast. It is a newspaper article.

The thesis is introduced with a short excursus to the history of using mathematical methods in linguistics and to the history of mathematical and quantitative linguistics themselves. An important part of the thesis is a developed algorithm of quantitative text processing which is supplemented with practical applications, calculation examples, tables, figures and graphs. The theory of language fractals is developed and the way of calculating the degree of semanticity is determined. The models of language fractals together with their associated mathematical fractals are visualized by means of the fractal theory. The

(non)suitability of quantitative exploration of poetical texts and setting up units is discussed in a comparative way.

1. Úvod

Pokud se týká Poeova *Havrana* a jeho překladů, Alena Dvořáková napsala v (Dvořáková, 2009) bez jakéhokoli dalšího vysvětlení, že „pátráme po hlubším významu někde ‚pod povrchem', v nevědomí básně: ve slabikách, jež musí být vypuštěny z tříslabičných slov, aby se při přednesu dodrželo metrum; *ve fraktálním rozložení slova Lenore* (jež z překladů mizí); v historickém kontextu, který má být ‚černým svědomím' básně."

Jak již bylo dříve zmíněno v (Andres, 2010), fraktály v poezii rozumíme obvykle sémantické rekurze. Byly též zmíněny dva konkrétní příklady objektů Vladimíra Holana a Wallace Stevense. Poznamenejme též, že současný maďarský básník Ferencz Gyõzõ (narozen v roce 1954) je autorem básně „Fraktální vědomí". Jak si povšiml Shannon v (Shannon, 1993), termíny v přirozených jazycích mohou být zdánlivě použity, aby byly popisovány stavy událostí na různých rozlišovacích úrovních.

Další typ fraktality byl detekován v (Becker & Flaxer, 2008) v tom smyslu, že organizace textu a neuronální aktivity mohou mít souvislost díky analogii mezi hierarchickou strukturou neuronální elektrické aktivity a hierarchií struktury textové. V (Henry, 1995) byl též heuristicky zvažován jazyk jako produkt mozku; bylo demonstrováno, že gramatika je svou povahou fraktální.

Žádný z těchto přístupů nebudu v této práci detailněji rozvíjet. Byly zmíněny, aby bylo zřejmé, že tato práce není první, ve které se zvažují fraktální vlastnosti jazyka nebo jeho produktů. Zde zvolený přístup vychází a zpracovává práce Luďka Hřebíčka na pozadí kvantitativní lingvistiky, viz (Hřebíček, 1997), (Hřebíček, 2002), (Hřebíček, 2007). Tento přístup byl dále rozvinut a formalizován v (Andres, 2009), (Andres, 2010), (Andres & Rypka, 2011) a v našich pracech (Andres et al., 2011) a (Andres & Benešová, 2011). Jan Andres též vytvořil teorii, jež umožňuje vizualizace jazykových fraktálů, o kterých bude v této práci řeč.

Vlastním cílem této práce je analyzovat textové výběrové soubory pomocí *Menzerath--Altmannova zákona* (MAL), viz (Altmann, 1980) a (Altmann et al., 1989). Metodologicky budu postupovat dle algoritmu, který byl zaveden v (Andres et al., 2011). Každý jednotlivý krok bude detailně popsán a opatřen konkrétními příklady i s ukázkami výpočtů. Navíc jsem výběrové soubory pro tento experiment vybrala tak, aby mi tento výběr umožňoval jejich srovnání s ohledem na stejné sémantické pozadí. Práce vznikla, aby na základě experimentů a analýz naznačila možnosti správných a efektivních způsobů segmentace textu na různých jazykových úrovních, aby vymezila způsoby testování MAL na různých textových výběrových souborech v různých jazycích, ale se stejným sémantickým základem, a způsoby testování výběrových souborů na fraktalitu.

Práce je na úvod v kapitole 2 a 3 doplněna krátkým přehledem historie koexistence matematiky a lingvistiky s přihlédnutím speciálně ke kvantitativní lingvistice a fraktální geometrii. Protože hlavní část práce zaměřuji na aplikaci fraktální geometrie v lingvistice, bylo nutné do třetí kapitoly integrovat sekci, která se věnuje této vědní disciplině. V další sekci je speciálně traktován Menzerath-Altmannův zákon a jeho využití pro dále popsanou analýzu. Hlavní část této práce, kapitola 4, je věnována popisu a detailům algoritmu fraktální analýzy výběrového souboru. Jedná se o algoritmus o devíti krocích. Každý z nich je popsán, jsou

komentovány problémy, které mohou nastat, a doplněny jsou ukázky výpočtů mechanických i pomocí počítačového softwaru. Výstupy výpočtů jsou vizualizovány, pokud je to možné. Je provedena shluková analýza na základě dat získaných analýzou fraktální. Poslední krok je zaměřen na interpretaci výsledků a výstupů. Na závěr, v kapitole 5 je výše zmíněný experiment doplněn tradičnějším zpracováním výběrových souborů pomocí teorie informace a numerické estetiky. V přílohách jsou shromážděny ukázky výpočtů, výstupy ve formě tabulek, obrázků či grafů. Pro velkou obsáhlost výstupů mohly být mnohdy uvedeny jen exemplární případy.

Pozoruhodné výsledky byly dosaženy též Benoitem Mandelbrotem v (Mandelbrot, 2003) a Alim Eftekharim v (Eftekhari, 2006). Vycházejí ze stejné interpretace, která ale není založena na MAL, ale na Mandelbrot-Zipfově zákonu. Iterační systémy funkcí, které hrají velkou roli v experimentu zde prezentovaném, jsou aplikovány v analýzách ve (Fernau & Staiger, 2001) a (Gutiérrez et al., 2003), na rozdíl od tohoto experimentu ale byly aplikovány na formální jazyky. Různé typy dalších přístupů a metod kvantitativní analýzy jsou dostupné ve (Wildgen, 2011) a (Wimmer et al., 2003).

Analýzám je podroben originální anglický text básně E. A. Poea *The Raven*, její překlad do německého jazyka a osmnáct různých překladů do jazyka českého, (Poe, 1931), (Poe, 1985), (Poe, 2008a), (Poe, 2008b). Jako nepoetický text byl zvolen článek ze Svitavského deníku (Nebeský, 2009).

2. Krátký historický náhled na používání matematiky pro lingvistické účely

Považuji za nutné a především účelné zahájit svou práci krátkým nástinem alespoň několika momentů, kdy se v lingvistice objevila jako nástroj na zkoumání jazyka a na vyhodnocování experimentů matematika. V raných obdobích, která zde budu zmiňovat, se spíše jedná o více či méně filosofické aktivity a úvahy jedinců, které měly více či méně zásadní vliv na rozvoj obou věd. Nešlo tedy zpočátku čistě o používání matematických metod v pravém slova smyslu. Výčet styčných bodů matematiky a lingvistiky nebude v žádném případě vyčerpávající, neboť toto není hlavní náplní této práce. Primárním účelem je navodit hned na začátku atmosféru interdisciplinarity mezi lingvistikou a matematikou. Jako svůj hlavní nástroj pro dosažení cíle této práci jsem si zvolila konkrétně jednu oblast matematiky, a to fraktální geometrii, přesto je ale signifikantní zmínit se o matematických metodách používaných v lingvistice obecně, aby bylo jasné, že to nebyl pouhý jeden případ sblížení dvou tak na první pohled odlišných věd.

Pomiňme naprosté počátky vývoje lidstva, kdy se přímo o vědeckém uvažování a exploraci v dnešním slova smyslu nedá příliš hovořit. Touto poznámkou nikterak nechci snižovat význam těchto období pro vývoj věd, spíše chci akcentovat sblížení obou věd v nejbližší minulosti a současnosti. Přenesme se přímo do tzv. předvědeckého období. V předvědeckém období se též nedá přímo mluvit o používání matematických metod v jazykovědě. Přesto však lingvistika, a samozřejmě nejen lingvistika, byla výrazně ovlivněna myšlením a prací mnoha matematiků, například *René Descarta* (1596 – 1650), významného filozofa a matematika. Mimo jiné je obzvlášť při zkoumání lidské mysli a idejí autorem teze, že plnohodnotné může být pouze to poznání, jež myslící rozvažování může vyjádřit ve zcela průzračných, rozumových, „matematických“ pojmech, viz (Störig, 1992). Jeho ideje a názory vedly k formování tzv. racionalismu.

Podobný vliv na lingvistiku mělo také několik dalších matematiků-filozofů, jmenujme alespoň *Gottfrieda Wilhelma Leibnize* (1646 – 1716). Leibniz se mimo jiné zabýval strukturováním jazyka. Ve svém katalogizování a abstrahování došel až téměř k tomu, co dnes nazýváme binárním kódem, tedy jazykem, kterým programujeme počítače. Šlo o jazyk umělý, který měl však zohledňovat popisované objekty a především měl být jednoduše srozumitelný a zapsatelný.

Stále se však jedná spíše o vliv jedince, nikoliv ucelené používání matematických myšlenek či metod. Větší změna měla přijít teprve se začátkem 20. století, kdy došlo v posunu náhledu na zkoumání jazyka. Jazyk přestal být zkoumán ve svém historickém vývoji (pohledem diachronním) a začal být analyzován jako entita existující v libovolné době nezávisle na historii (synchronní lingvistika). Na počátku 20. století se začala v lingvistice hojně využívat logika a predikátová logika, jež je založena na pravidlech dedukce a závisí na určitých lingvistických strukturách.

Zásadní událostí pro vývoj lingvistiky nejen ve vztahu k matematice se stalo vydání „Kurzu obecné lingvistiky“ (1916) *Ferdinanda de Saussura*. Zejména de Saussurova teorie jazykového znaku jako základního stavebního kamene nutně přispěla k exaktnějšímu náhledu na jazyk. Jazyk je považován za systém, jehož prvky jsou navzájem spolu spjaty určitými vztahy,

a právě tyto vztahy mají být pomocí strukturní komparativní analýzy synchronicky (v jedné časové rovině) zkoumány. Jazyk jako systém znaků, který slouží k přenosu informací, se stal předmětem výzkumu dále např. *strukturalistů* (škola ženevská, kodaňská, pražská).

Představitel tzv. kodaňské školy *Viggo Bröndal* (1887 – 1942) se zasloužil o zapojení používání aparátu a metod symbolické logiky v jazykovědě. Také pravděpodobně nejvýznamnější člen této školy *Luis Hjelmslev* (1899 – 1965) prosazoval používání logických a matematických symbolů při zkoumání jazyka. Dále se v jeho díle objevuje pojem funkce, jehož pojetí se velice blíží pojetí funkce v matematice.

Z hlediska používání matematiky pro účely jazykovědy je též důležité zmínit sovětskou školu, zejména kvůli působení v oblasti algebraické a strojové lingvistiky. V šedesátých letech se zde také zrodil tzv. aplikačně-generativní model gramatiky, který využívá logických a matematických metod.

Zakladatel deskriptivismu *Leonard Bloomfield* (1887 – 1949) klade důraz na používání vědeckých postupů v lingvistice v souladu tzv. logickým pozitivismem či novopozitivismem, což je filosofický směr ovlivněný pracemi z oblasti logiky a matematiky – pokus vyjádřit všechna smysluplná tvrzení kombinacemi výrokové logiky a smyslových vjemů. Na Leonarda Bloomfielda navázal s pomocí formální logiky a matematiky *Zellig S. Harris* (1909 – 1992) ve svých „Metodách strukturální jazykovědy“. Jeho snahou bylo najít vhodný popis jazyka exaktními prostředky bez ohledu na význam příslušných jazykových jednotek, čímž se jeho úsilí řadí opět mezi předchůdce moderní algebraické lingvistiky.

Na strukturalismus v Americe navazuje deskriptivní a generativní mluvnice se svým nejvýznamnějším představitelem *Noamem Chomskym* (nar. 1928). Prof Chomsky ve svém díle usiluje o popis a zkoumání neviditelných, abstraktních jazykových struktur, které vytvářejí ze slov správnou větu modelováním pomocí matematiky. Matematiku považuje za nejpřesnější nástroj k popisu abstraktní struktury. A jazyková struktura bezpochyby je vysoce abstraktní strukturou. Gramatika jazyka je pak množinou axiomů. Opět se tedy jedná o algebraickou lingvistiku.

Pro další detaily a informace viz např. (Černý, 1996) a (Struik, 1963).

50. a 60. léta 20. století se již vyznačují nástupem matematické lingvistiky v pravém slova smyslu, proto tomuto odvětví věnuji samostatný prostor.

3. Matematická lingvistika, pojem a jeho obsah

When you can measure what you are speaking about,
and express it in numbers, you know something about it.
When you cannot measure it, when you cannot express it in numbers,
your knowledge is of a meager and unsatisfactory kind;
It may be the beginning of knowledge,
But you have scarcely, in your thoughts, advanced to the stage of science.

Lord Kelvin

Předmětem zájmu a působení oboru matematická lingvistika je popis jazyka matematickými metodami. Může jít konkrétně o metody symbolické (algebraické, formální) nebo o využití statistiky, případně kombinace těchto metod. Další oblastí matematické lingvistiky je lingvistika korpusová, která se zabývá přípravou a využitím elektronických textových korpusů. Velice důležitou funkcí, kterou matematika může přispět k rozvoji lingvistiky, je verifikace či vyvracení rozličných nastolených hypotéz.

Jako oficiální počátek matematické lingvistiky se někdy uvádí VIII. Mezinárodní lingvistický kongres v Oslu v roce 1957. V současné době pod pojmem matematická lingvistika můžeme vidět zejména čtyři samostatně se rozvíjející pole působnosti. Jsou to: kvantitativní lingvistika (někdy označovaná pro své hojné využívání statistických metod jako statistická lingvistika, nejde však pouze o statistiku, která poskytuje v tomto úhlu pohledu prostředky ke zkoumání jazyka, jak bych dále chtěla ukázat), algebraická a strojová lingvistika. K nim přibývá již výše zmíněná korpusová lingvistika. Protože by se tato práce měla úzce týkat kvantitativní lingvistiky, pro úplnost bych ráda podiskutovala nejdříve v krátkosti algebraickou a strojovou lingvistiku.

Algebraická lingvistika, viz např. (Černý, 1996) a (Sgall et al., 1974), nepoužívá při zkoumání jazyka metody statistické. Její počátky již můžeme najít v druhé polovině 19. století, kdy se prudce rozvíjela formální logika a přinášela s sebou abstraktní a nekvantitativní postupy. Název po tuto vědní disciplínu byl navržen *Y. Bar-Hillelem* v druhé polovině 50. let 20. století. Algebraická lingvistika se zabývá rozborem uspořádaných řetězců jazykových jednotek převážně na syntaktické úrovni jazyka. Mezi nejznámější a nejdůležitější patří již zmíněná generativní mluvnice *Noama Chomského*, funkční generativní popis *P. Sgalla*, aplikačně generativní model jazyka *S.K. Šaumjana*, rekognoskativní a kategoriální gramatika, analytické modely jazyka, závislostní gramatika a další.

Ve druhé polovině 20. století dochází také k raketovému rozvoji výpočetní techniky, který byl bezprostředně způsoben prudkým nárůstem objemu informací, které bylo potřeba zpracovat či pouze uložit. S tím jsou také spojeny zvýšené nároky na překladatele a rychlost překládání. Spojením a praktickou aplikací teoretické kvantitativní a algebraické lingvistiky vzniká tzv. *strojová (počítačová) lingvistika*, anglický původní název je computational linguistics. Při práci s počítači jsou opět využívány metody matematické lingvistiky. Vzniklá nutnost najít způsob strojového a tedy rychlejšího překladu textů pomocí počítačové techniky se ale bohužel ukázala do značné míry neřešitelnou. Problém se objevil v sémantické složce

přirozených jazyků, která se vzpírá konkrétnímu popisu a tudíž i převodu do jiného přirozeného jazyka pomocí počítače. Uplatnění ve strojové lingvistice mají ale i jiné činnosti, jako například spektrální analýza mluvené řeči, zpracování frekvenčních seznamů apod. Při jejich zpracovávání se spoléháme především na operační rychlost počítačů. Další informace viz např. (Hajičová et al., 2002).

Kvantitativní lingvistiku definuje *Marie Těšitelová* následujícím způsobem: „Kvantitativní lingvistika je složka matematické lingvistiky, která kvantifikuje (zjišťuje kvantitativní data) jevy různých jazykových rovin a modeluje jejich vztahy realizující se ve větě, v textu, abychom lépe poznali jejich příčinný mechanismus, jejich fungování, jejich stránku formální, ale i sémantickou. Vzhledem k tomu, že se při aplikaci kvantitativních metod v lingvistice zatím v převážné míře užívá statistiky, mluví se též někdy o lingvistice statistické. Je to ovšem termín užší než termín kvantitativní lingvistika. Nelze je dobře ve všech případech zaměňovat," viz (Těšitelová, 1987, str. 8-9).

Třetí odvětví matematické lingvistiky má své kořeny, pokud jde o uplatnění kvantitativních metod, už v díle *Jana Amose Komenského* „Janua linguarum reserata", kde tento učenec ukázal, jak je možno využít znalostí o frekvencích slov k výuce cizího jazyka. Mezi dalšími, kteří už v půli 19. století upozorňovali na možnost využití kvantitativních metod v lingvistice, patřil bezesporu i ruský matematik a jeden z nejvýznamnějších členů ruské matematické školy *V. J. Buňakovskij*. Na konci téhož století mladogramatik *Herman Paul* tvrdí, že jazyk je statistickým průměrem jazykových projevů všech jeho uživatelů, dále používá pojem invariantní hlásky ve fonetice. V osmdesátých letech stejného století aplikoval matematik *August Seydler* ve svém „Počtu pravděpodobnosti v přítomném sporu" nástroje pravděpodobnosti při rozhodování o pravosti tzv. „Rukopisů královédvorského a zelenohorského". Takřka o sto let později se v roce 1962 snažili američtí matematici *Frederick Mosteller* a *David Wallace* najít v jazyce textů v „Listech federalistů" struktury schopné určit autora. V sedmdesátých letech 19. století se jako vůbec první lingvista *William D. Whitney* zabýval frekvencí (anglických) hlásek. **Frekvence** neboli **četnost výskytu** je jeden z nejdůležitějších pojmů kvantitativní lingvistiky důležitý nejen po samotnou jazykovědu, ale také například pro metodiku výuky jazyků (výběr nejdůležitější slovní zásoby pro studenty cizích jazyků různých úrovní), donedávna také pro tiskaře, stenografy (nejfrekventovanější slova mají nejjednodušší symboly), pro tvorbu a výrobu nejrůznějších her a hlavolamů apod. Také *Samuel Morse* při sestavování znaků pro svou abecedu využíval tyto poznatky a pro nejfrekventovanější písmeno v anglické abecedě vybral nejjednodušší znak atd. První **slovník četnosti** se objevil na samém konci 19. století. Sestavil jej německý stenograf *F. W. Käding* a nazval „Slovník četnosti výskytu německého jazyka". K zajímavým a důležitým zjištěním vyplývajícím ze znalosti četnosti výskytů slov v jazycích se vrátím později. Tyto a další milníky viz např. (Černý, 1996), (Těšitelová, 1987a), (Těšitelová, 1987b) a (Devlin, 2002).

Na počátku 20. století přispěl k rozvoji kvantitativní lingvistiky velice významně ruský matematik *Andrej A. Markov*, když ve svém díle „Příklad statistického výzkumu textu Evžena Oněgina" dospěl k závěru, že v každé části výpovědi lze s určitou pravděpodobností předvídat, které jazykové jednotky budou následovat. Množství informace přenášené jazykovou jednotkou se tedy dá měřit. Tato významná teorie dostala název **Markovův proces**. Mluvení je

podle Markova proces, ve kterém jsou k jednotlivým jazykovým jednotkám už vysloveným připojovány další podle relativní frekvence, která je pro daný jazyk závazná. Zákonitosti se týkají jednak frekvence písmen (a také mezer), pravděpodobností, že po nějakém písmenu následuje další, a faktu, že různá písmena (stejně tak i slabiky a slova) nesou různou **míru informace**. Tyto poznatky úzce souvisí s tzv. teorií informace. Podle ní největší množství informace nesou vždy jednotky předem nejobtížněji odhadnutelné (viz. entropie). Redundantní je potom taková jednotka, s jejíž existencí předem počítáme. Pravděpodobnost výskytu jazykové jednotky je přímo úměrná frekvenci dané jednotky v jazyce, více viz (Bartók & Janoušek, 1980) a kapitola *Havran* a teorie informace.

Ve dvacátých a třicátých letech se do historie rozvoje kvantitativní lingvistiky zapsal americký lingvista německého původu, profesor harvardské univerzity *George Kingsley Zipf*. Zabýval se studiem relativní frekvence hlásek a došel ke zjištěním, že hlásky a jejich třídy v různých textech jazyka mají stejnou frekvenci. Ve všech jazycích je počet neznělých hlásek přibližně dvakrát větší než počet znělých. A čím je obtížnější hlásky z hlediska jejich artikulace produkovat, tím mají menší frekvenci, což souvisí s principem ekonomie v jazyce, viz (Těšitelová, 1987a). Podobnými problémy se zabýval již koncem 19. století francouzský psycholog *B. Bourdon* a Francouz *J. B. Estoup*.

První Zipfův zákon:

V jazyce působí dvě protikladné síly, sjednocující a rozlišující, snažící se, aby v jazyce měla slova co největší frekvenci, a tím jich bylo co nejméně, a zároveň aby jazyk disponoval s co největším počtem slov majících nízké frekvence.

$$r \cdot f = k \tag{1}$$

r rank slova (či jeho pořadí)

f absolutní frekvence příslušného slova

k konstanta

Čili čím je rank daného slova nižší, tím je jeho frekvence vyšší. Mezi rankem a absolutní frekvencí platí nepřímé úměrnosti.

Francouzský matematik *Benoit Mandelbrot*, který má podstatnou zásluhu na zkoumání a popularizaci **teorie fraktálů** a **fraktální geometrie**, o které bude řeč později, ukázal, že Zipfův vzorec sice udává obecný spád křivek, ale velmi špatně zobrazuje podobnosti. Přidal tedy své úpravy prvního zákona:

Harmonický zákon:

$$p_r = \frac{P}{r} \tag{2}$$

p_r příslušná četnost

r rank

P konstanta po každý text

Kanonický zákon (závisí na počtu slov, která máme k dispozici)**:**

$$p_r = P(r+\rho)^{-B} \tag{3}$$

neboli $\log p_r = \log P - B \cdot \log(r+\rho)$

P, B, ρ konstanty, parametry textu.

Harmonický zákon je pouze zvláštním případem kanonického, platí, jestliže $B = 1$ a $\rho = 0$. Tento zákon dále *Mandelbrot* rozvinul v takzvané lexikografické stromy (viz. dále v textu). *Marie Těšitelová* dále zjistila spolehlivost zákona pro slova, jejichž rank se kryje s pořadím. Pro slova s vysokou, nebo naopak nízkou frekvencí zákon vystihuje již vztahy hůře, viz (Těšitelová, 1987a).

Druhý Zipfův zákon:

Druhý Zipfův zákon vyjadřuje vztah mezi frekvencí slova a počtem různých slov, které tuto frekvenci mají. Čím je frekvence nižší, tím více slov tuto frekvenci má.

$$a \cdot b^2 = k \tag{4}$$

a počet slov s jistou frekvencí

b frekvence těchto slov

k konstanta

Zipf předpokládá, že tento zákon platí pro všechny jazyky, avšak vylučuje z něj slova s nejvyšší a nejnižší frekvencí. Formule ale také neplatí stejně pro texty různé délky, viz (Těšitelová, 1987a).

Třetí Zipfův zákon:

Vyjadřuje vztah mezi frekvencí slova a počtem jeho významů. Jeho závěrem je, že počet různých významů (polysémie) je vyšší u slov s vyšší frekvencí.

$$\frac{m}{\sqrt{f}} = k \tag{5}$$

m počet významů daného slova

f frekvence tohoto slova

k konstanta

Marie Těšitelová však prokázala, že tento vztah platí většinově jen po slova formální. Z tohoto pak dále vyplývá, že čím je slovo delší, tím má nižší frekvenci, což v zásadě platí pro všechny jazyky s ohledem na jejich typologii, viz (Těšitelová, 1987a).

Vývoj kvantitativní lingvistiky v našich zemích, mimo již výše zmíněných, je naprosto neodmyslitelně spojen s pracemi lingvistů, jako byli např. *Vilém Mathesius*, *Bohumil Trnka* či *Josef Vachek*. Jejich díla byla spojena především s oblastí fonologickou a lexikální. V neposlední řadě zde také vznikl za přispění pedagoga *Václava Příhody* a bohemisty *Vladimíra Šmilauera* český frekvenční slovník autorů *J. Jelínka*, *J. V. Bečky* a *M. Těšitelové*, který byl ve svých začátcích obzvlášť míněn po účely pedagogicko-metodické.

Závěry vyplývající z frekvenčních slovníků a dalších děl, které statisticky zpracovávají přirozené jazyky, jsou velice zajímavé a důležité po různá vědní odvětví, proto považuji za důležité zmínit alespoň některé z nich s odvoláním obzvláště na práci *Marie Těšitelové*, viz např. (Těšitelová, 1987a), (Těšitelová, 1987b). Slova ve zmíněných frekvenčních slovnících jsou seřazena do tří úrovní: slova s nejvyšší a vyšší (prvních deset slov), slova se střední a konečně s nízkou a nejnižší frekvencí. Rozložení těchto tří kategorií se blíží exponenciálnímu rozložení, podobně jako se tomu děje u mnoha dalších jevů v přírodě. Slova s nejvyšší frekvencí jsou až na jednu výjimku (desáté pořadí má slovo *který*, které se z ekonomických důvodů v hovorovém jazyce stejně zkracuje na *co*) jednoslabičná – princip ekonomie v jazyce. Těchto deset slov pokrývá ve většině jazyků průměrně dvacet procent textu (první slovo přibližně pět procent, desáté jedno procento textu), což je důležité pro výuku cizích jazyků, stejně tak jako např. pro dešifrování zakódovaného textu. Většinou se jedná o slova formální, nebo ta, která poklesají dokonce na částice (slova vycpávková – fillers), odtud jejich vysoká frekvence obzvláště v mluveném projevu.

Slova ze střední kategorie jsou vymezena svou horní a dolní mezí, tedy slovy s nejvyšší a nejnižší frekvencí. Bývají to zpravidla méně frekventovaná slova gramatická, adverbia zejména zájmenného původu, substantiva a adjektiva – tedy většinou slova plnovýznamová. Rozsah tohoto pásma závisí na rozsahu korpusu, materiálu, funkčním stylu, slohových útvarech apod.

Třetí kategorie jsou slova s nejnižší frekvencí 10 - 1, což jsou nejčastěji slova plnovýznamová. Texty umělecké mají slov s frekvencí 1 a 2 více než texty stylu věcného. Lexikální jednotky s frekvencí jedna až deset určují tzv. **bohatství slovníku**, naopak nejfrekventovanější jednotky definují **koncentraci slovníku**. Pro bohatství a koncentraci slovníky se pokusil *P. Guiraud* vytvořit v lexikální statistice dvě formule:

Bohatství slovníku:

$$\text{pro všechna slova obecně} \qquad R = \frac{V}{\sqrt{N}} \tag{6}$$

$$\text{nebo pouze pro plnovýznamová slova} \quad R = \frac{V}{\sqrt{2N}}\text{, kde} \tag{7}$$

R bohatství slovníku

V slovník, tj. počet všech lexikálních jednotek, lexémů, různých slov (FSČ V=54 486)

N délka textu, celkový počet slov (FSČ N=1 623 527)

Bohužel však je opět dokázáno, že tato formule neplatí pro češtinu, ani pro flexivní jazyky obecně, platí pouze pro analytické jazyky, viz (Černý, 1996).

Koncentrace slovníku:

$$C = \frac{\sum_{50}^{1}}{N} \tag{8}$$

C koncentrace slovníku

Tzn. koncentrace je vyjádřena jako poměr prvních padesáti nejfrekventovanějších slov ku délce textu.

Avšak *Marie Těšitelová* dospěla při svých výzkumech k závěru, že při stanovení bohatství slovníku z hlediska kvantitativního je potřeba vzít v úvahu tři hlediska:

1) **Rozsah slovníku** (v češtině je třeba počítat pouze s 80 % textu – se slovy plnovýznamovými, event. se 70 %, bereme-li v úvahu za slova plnovýznamová pouze substantiva, adjektiva, slovesa a adverbia, a ne zájmena a číslovky)

$$R = 100 \cdot \frac{V}{\frac{8}{10} \cdot N}, \text{ event. } R = 100 \cdot \frac{V}{\frac{7}{10} \cdot N} \quad . \tag{9), (10}$$

2) **Rozptýlení slovníku**, které ukazuje specifiku jazyka stylu funkčního, ale i individuálního.

$$D = 100 \cdot \frac{V_1^{10}}{V} \tag{11}$$

V_1^{10} počet plnovýznamových slov s frekvencí 1-10.

3) **Koncentrace slovníku** ukazuje, jaký podíl slovníku textu, popř. autora, připadá na slova nejfrekventovanější.

$$K = 100 \cdot \frac{N_1^{10}}{N} \tag{12}$$

N_1^{10} délka textu odpovídající prvním deseti nejfrekventovanějším slovům.

Mezi dalšími českými lingvisty působícími v kontaktu s kvantitativní lingvistikou můžeme dále jmenovat *Jiřího Krámského*, žáka V. Mathesia a B. Trnky, který se zabýval fonologickým rozborem hlásek, vzájemnými vztahy mezi fonémy v různých pozicích ve slově a typologií jazyků, zajímal se o orientalistiku, anglistiku a o metodiku vyučování cizích jazyků, ve svých pracech užíval kvantitativní metody.

Dále jmenujme např. *Ladislava Nebeského*, docenta pro obor matematika, směr algebra a teorie čísel, který se zaměřuje na matematiku pro lingvistiku, konkrétně pro fonetiku na Univerzitě Karlově v Praze.

Mezi další velice významné české lingvisty, kteří se věnují kvantitativní lingvistice, tentokrát můžeme říci důsledně kvantitativní, nikoli statistické lingvistice, patří bezesporu orientalista *Luděk Hřebíček*, který se zabývá dokazováním vztahu jazyka na všech jeho subsystémech k významu pomocí fraktální struktury jazyka v souvislosti s textovou lingvistikou. Tomuto tématu bych se dále chtěla věnovat později po tomto všeobecném úvodu, viz (Hřebíček, 1997), (Hřebíček, 2002).

3.1 Úvod do kvantitativní analýzy Altmannovsko-Hřebíčkovského typu s aplikací teorie fraktálů

„Veda, ktorej chýbajú hypotézy, je protoveda a veda, ktorej hypotézy sú netestovateľné, je pseudoveda…
Vo filologických vedách existujú dodnes poddisciplíny, ktoré sa uspokoja s tým,
že rozmnožujú batériu pojmov, vytvárajú množstvo ‚-izmov' a ‚-ém' na opis
a klasifikáciu javov a žijú v domnienke, že vytvárajú teóriu…"

Gabriel Altmann, viz (Wimmer et al., 2003)

Za dva zásadní milníky 20. století považuje Luděk Hřebíček vydání de Saussurova „Kurzu obecné lingvistiky" a přínos prof. Gabriela Altmanna. Dokonce píše, že „skutečná lingvistika druhé poloviny 20. století je altmannovskou lingvistikou," viz (Hřebíček, 2008). Jak bylo zdůrazněno výše, Gabriel Altmann klade důraz na formulaci vědeckých zákonů neboli testovatelných hypotéz. A hypotézy je třeba přijímat či zamítat, a to nejlépe pomocí statistických metod. Prof. Altmann je bezpochyby jeden ze zakladatelů moderní kvantitativní lingvistiky.

V souvislosti s kvantitativní lingvistikou je též nutné zmínit tři periodika publikující výsledky nejnovějších výzkumů na tomto poli. Prvním z nich je „Journal of Quantitative Linguistics" (Official Journal of the International Quantitative Linguistics Association). Vydavatelem Journal of Quantitative Linguistics je prof. *Reinhard Köhler* z Trevíru, vůdčí osobnost Oddělení pro zpracování lingvistických dat na tamní univerzitě. Jeho koeditorem je prof. *Gabriel Altmann*, nesporně světová autorita v kvantitativní lingvistice, původem slovenský jazykovědec, dnes uznávaný zakladatel tohoto oboru v Německu. O náplni časopisu píše Luděk Hřebíček následující: „Mají-li vědy za úkol přinášet explanaci předmětu poznání ve formě odmítnutelných (testovatelných) teorií, neexistuje důvod, proč by lingvistická teorie měla mít jiné cíle. Při plnění tohoto úkolu se neobejde bez kvantitativního přístupu, ačkoliv ten zajisté není nezbytnou podmínkou, pokud teorie je schopna splnit požadavek odmítnutelnosti jinak. Zatím ovšem věc vypadá tak, že to jinak nejde. Valná část toho, co se označuje pojmem lingvistika, je v podstatě hledání jakéhosi návodu k tvoření správných věd. To je účelné a rozumné, je to praktické a potřebné, není to však vědecké, pokud věda představuje nerozporné soustavy teorií. Přesvědčivým dokladem toho je fakt, že dnes převládající teorie jazyka zcela ztroskotávají, když mají přejít k explanaci nadvětných útvarů. Při pozornějším pohledu je zřejmé, že klasická lingvistická teorie kromě učených pojmenování v interpretativních výrocích, nemajících většinou povahu vědecké teorie, nenabízí nějakou podstatnou informaci o povaze jazyka," viz (Hřebíček, 1994). Druhým, neméně významným, je „Glottometrics", který je vydáván pod vedením prof. Dr *Gabriela Altmanna*. Třetím časopisem

je „Glottotheory“, který je vydáván na Filozofické fakultě Univerzity sv. Cyrila a Metoda ve slovenské Trnavě a jehož editorem je Emília Nemcová. Zaměřen je na metodologické a teoretické problémy jazyka a textu a empirickou charakterizaci jazyka a textu kvalitativním a kvantitativním způsobem.

V experimentu, který bude v následujících kapitolách této práce popsán, vycházíme z de Saussurova učení a teorie o principu linearity, viz (Andres, 2009). Navzdora mnohým kontradikcím je tato teorie chápána ve smyslu Hřebíčka a mnoha dalších kvantitativních lingvistů, viz (Hřebíček, 1995), (Wimmer et al., 2003), kteří považují text v de Saussurových liniích za lineární nástroj pro transfer nelineárního chápání a rozpoznávání, poněvadž vzniká z multidimenzionálních znalostí vyslovených jednodimenzionálním způsobem. V tomto smyslu je možné rozlišovat šest typů linearizací: mentální, kontextuální, gramatickou, poetickou, stochastickou a chaotickou, viz dále (Andres, 2009) a (Wimmer et al., 2003).

Na de Saussurovo pojetí navázal v našem slova smyslu Luděk Hřebíček a uchopil jej v exaktní matematické podobě, viz (Hřebíček, 1997), (Hřebíček, 2002), (Hřebíček, 2007), (Wimmer et al., 2003) a dále v textu. Hřebíčkova heuristická teorie byla poté formalizována prof. Janem Andresem v (Andres, 2009) a (Andres, 2010). Na základě této koncepce byla hypotéza testována a prováděny experimenty, které jsou prezentovány a komentovány v této práci a v (Andres et al., 2011) a (Andres & Benešová, 2011). Zároveň byla vybudována systematizovaná metodologie a příslušný aparát, viz dále v textu a v (Andres et al., 2011).

3.1.1 Co je to fraktál

Před zahájením pojednání o experimentu samotném, jedním z jehož nástrojů je teorie fraktálů, považuji za nutné předložit velice hrubý nástin této teorie, stručnou historii jejího vzniku, nejpopulárnější příklady a pro heuristickou ilustraci použití mimo lingvistiku samotnou.

Co je to fraktál? Hned první otázka přináší značné problémy, neboť neexistuje jednotná definice fraktálu. Existuje několik přístupů k definici fraktálu, které ale vzájemně nekoincidují. Jmenujme tři základní matematické přístupy k nadefinování objektu zvaného fraktál, více viz (Andres, 2010):

Definice 1. Říkáme, že množina F_1 je *fraktálem ve smyslu Mandelbrota* (psáno $F_1 \in \mathcal{F}_1$), jestliže jeho fraktální dimenze není celé číslo.

Existuje několik definic fraktální dimenze, např. soběpodobnostní, Hausdorff-Besicovitchova, viz např. (Falconer, 1990).

Definice 2. Říkáme, že množina F_2 je *fraktál ve smyslu Hutchinson-Barnsleyho* (psáno $F_2 \in \mathcal{F}_2$), jestliže existuje (konečný) systém kontrakcí $\{T_i: X \to X | i = 1, \dots, n\}$ na úplném metrickém prostoru (X, d), který se nazývá *IFS* (iterated function system), takový, že

$$\cup_{x \in F_2} \cup_{i=1}^{n} T_i(x) = F_2.$$

Zobrazení

$$\bigcup_{i=1}^{n} T_i: X \to 2^x \backslash \{\emptyset\}$$

se nazývá *Hutchinson-Barnsleyovo zobrazení*.

Fraktály jsou dle tohoto pojetí považovány za invariantní množiny (tzv. atraktory) s danými vlastnostmi.

Definice 3. Říkáme, že množina F_3 je *fraktál v axiomatickém smyslu* (psáno $F_3 \in \mathcal{F}_3$), jestliže vykazuje nekonečně se opakující *soběpodobnost* (tj. invarianci vůči měřítku).

Opět je nutné poznamenat, že existuje několik typů soběpodobnosti, např. matematická, kvasi, statistická, náhodná, stochastická. *Soběpodobnostní dimenzi* pak můžeme aplikovat na soběpodobné struktury následujícím způsobem:

$$D = \frac{\ln N}{\ln(1/r)}, \tag{13}$$

kde *D* je soběpodobnostní dimenze, *N* je celková délka útvaru v jeho částech a *r* je faktor zmenšení, viz dále v textu.

Definice 1, 2, 3 nemusí nutně korespondovat, viz (Andres, 2010).

Třetí z definic fraktálu, nejvíce heuristická a nejčastěji citovaná, tedy říká, že *fraktál* je geometrický (a z dalšího budiž zřejmé, že pravděpodobně nejen geometrický) objekt, který vykazuje tvarovou podobnost se svými částmi, tuto vlastnost nazýváme *soběpodobnost*. Pro jednodušší představu si v této souvislosti připomeňme list kapradiny nebo například hlávku brokolice. Každá větvička kapradiny se svým tvarem a strukturou podobá celému listu, viz obr. č. 1. Pokud uvažujeme objekty matematicky konstruované, mluvíme o tzv. *striktních* nebo *matematických fraktálech*, pro které je charakteristická stoprocentní a nekonečná soběpodobnost, viz např. (Hřebíček, 2002). V přírodě takové objekty nenajdeme. Ale matematika by přece měla „jen“ co nejvěrněji modelovat přírodu a reálný svět. Připomeňme Kennetha Falconera ve (Falconer, 1990): „V přírodě neexistují skutečné fraktály. (Dokonce tam neexistují skutečné přímky nebo kružnice!)“ Reálně existující objekty jsou tedy „pouhými“ aproximacemi matematických fraktálů v tomto smyslu definice.

Obr. č. 1: Počítačem vygenerovaná kapadina

Zkoumání fraktálů se věnuje tzv. *fraktální geometrie*, jejímž pravděpodobně nejznámějším představitelem je francouzský matematik *Benoit Mandelbrot*, který je nazýván

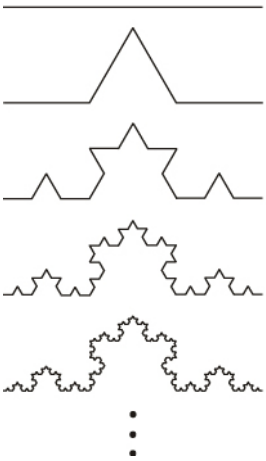
jejím zakladatelem. Mandelbrot v roce 1975 zavedl a poprvé ve své knize „Les objets fractals, forme, hasard et dimension“, (Mandelbrot, 2003), použil termín fraktál[1].

Uveďme si pro názornost několik příkladů nejznámějších matematických (v důsledku teoreticky stoprocentně soběpodobných) a přírodních objektů, které vykazují fraktální vlastnosti, a výpočtu jejich soběpodobnostní dimenze. Prvním typem budou „uměle“ vygenerované fraktální objekty, jejichž struktura bude hluboká ad infinitum a druhá kategorie budou příklady fraktálů, jejichž aproximace je možno najít v přírodě. Pro ilustraci první kategorie bych ráda uvedla dva fraktály, které bezpochyby patří mezi nejčastěji zmiňované. *Cantorova množina* vychází z úsečky, ze které v první iteraci vyjmeme střední, jednu třetinu původní délky dlouhou úsečku, stejně pak pokračujeme ve druhé iteraci se zbylými dvěma úsečkami atd. do nekonečna, obr. č. 2. Tento fraktál není vybrán samoúčelně. Tento typ konstrukce byl vybrán pro vizualizaci jazykových fraktálů v jednom z kroků algoritmu fraktální analýzy textu, který bude prezentován dále. Vybrán byl proto, že vycházíme při konstrukci z úsečky, tedy lineárního objektu, že tedy na první pohled heuristicky připomíná strukturu textu. Tento fakt na první pohled koresponduje s de Saussurovou teorií o lineární povaze označujícího, jež je povahou auditivní a jehož vzorek je měřitelný v jedné dimenzi, viz (de Saussure, 2007). Během konstrukce jednotlivé elementy původního elementu z iterace 0. mizí dle zadaného algoritmu, na rozdíl od druhého vybraného fraktálu.

Obr. č. 2: Cantorova množina

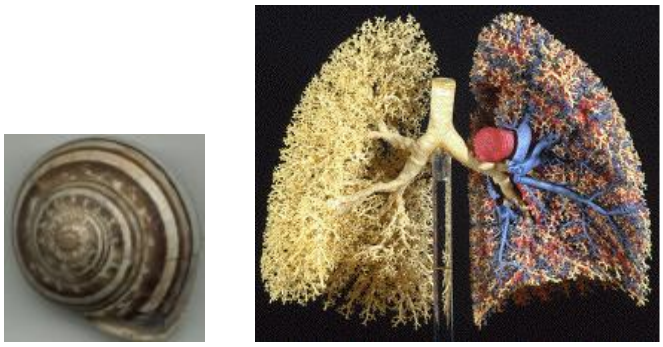
Tím příkladem je *Kochova křivka*, obr. č. 3. V tomto typu konstrukce s každou iterací, na rozdíl od Cantorovy množiny, elementy v objektu přibývají. I konstrukce může být použita pro vizualizaci jazykových fraktálů, viz (Andres & Rypka, 2011).

[1] Slovo vzniklo z latinského fractus, coz znamená nepravidelný, polámaný.



Obr. č. 3: Kochova křivka

Příklady „přírodních“ fraktálů[2], obr. č. 4, jsou například dokonalé sítě žil a tepen (jež zabírají minimum místa, a přesto nelze odebrat ani miligram lidského masa, aniž by byla prolita krev), bronchiální větvení, vylučovací soustava, ale i struktury ulit plžů, list kapradiny, blesky apod.



Obr. č. 4: Příklady fraktálních objektů v přírodě

„Klasické“ fraktály jako např. Cantorova množina, Sierpinského trojúhelník a Kochova křivka obvykle splňují všechny tři na úvod vyslovené definice fraktálu, viz (Andres, 2009).

V tab. č. 1 a 2 jsou pro ilustraci demonstrovány výstupy výpočtů fraktálních dimenzí některých z výše uvedených fraktálních objektů.

[2] To jsou takové objekty, jejichž struktura není a z důvodů reálné existence ani nemůže být nekonečná, ale velice často je možné najít matematický fraktál (model), jehož aproximací je daná přírodní struktura. To bude i mou snahou v dále popisovaném experimentu, tzn., najít matematický model, jehož aproximací by byla daná struktura textového výběrového souboru, a který bych byla schopna kvantitativně popsat.

Objekt	Počet částí	Faktor zmenšení	Dimenze
úsečka	např.3	$\frac{1}{3}$	1
čtverec	např. $6^2 = 36$	$\frac{1}{6}$	2
krychle	např. $6^3 = 216$	$\frac{1}{6}$	3
Cantorova množina	2^k	$\frac{1}{3^k}$	$\frac{\log 2}{\log 3} = 0{,}6309$
Sierpinskeho trojúhelník	3^k	$\frac{1}{2^k}$	$\frac{\log 3}{\log 2} = 1{,}5850$
Kochova křivka	4^k	$\frac{1}{3^k}$	$\frac{\log 4}{\log 3} = 1{,}2618$

Tab. č. 1: Příklady různých matematických fraktálních objektů a jejich fraktálních dimenzí

Přírodní objekt	Odhad fraktální dimenze
pobřeží	1,26
povrch mozku člověka	2,76
neerodované skály	2,2 – 2,3
obvod 2D-průmětu oblaku	1,33

Tab. č. 2: Příklady různých přírodních objektů a jejich fraktálních dimenzí

Podobně jako výše uvedené fraktální struktury objevující se v přírodě vykazují soběpodobnost na omezeném počtu úrovní, je možné najít a dokázat fraktální struktury a vlastnosti v lingvistických strukturách. V lingvistice je ale situace mnohem složitější, neboť je analýze podroben abstraktní objekt.

3.1.2 Menzerath-Altmannův zákon

Již před dlouhými časy lingvistika pochopila, že jazyk je živoucí organismus, systém složený z jednotek existujících na různých jazykových úrovních. Pokud byly tyto jazykové úrovně podrobeny běžné, po staletí praktikované analýze, byly od sebe odtrženy a jejich jednotky byly extrahovány, aby tak byly prozkoumány vztahy mezi nimi v rámci jednotlivých vět. Dlouhý čas analýza fungovala tímto způsobem, ale zásadní problémy nastaly s příchodem textové lingvistiky, která se pokusila zkoumat také nadvětné jazykové úrovně. Jako světlo v temnotě se poté zjevil Menzerath-Altmannův zákon.

V roce 1928 popsal Paul Menzerath vztah mezi délkou slova ve slabikách a délkou slabik ve fonémech. Tento vztah může být vyjádřen následujícím způsobem: čím delší je slovo, tím kratší je průměrná délka jeho slabiky, viz (Altmann, 1980). Vztah byl později zobecněn a formulován pomocí matematického vzorce profesorem Gabrielem Altmannem a je nazýván na počest obou vědců Menzerath-Altmannův zákon (MAL). Ve své komplexnější a obecnější podobě, která pokrývá a spojuje všechny známé jazykové úrovně, upřesňuje vztah mezi náhodně zvolenou jazykovou jednotkou na vyšší jazykové úrovni (*konstrukt*) a jejím

konstituentem/konstituenty na nejbližší nižší úrovni (*konstituent*). Slovní formulace *Menzerath-Altmannova zákona* (MAL) říká, že *čím delší je jazykový konstrukt, tím kratší jsou jeho konstituenty*. Zkrácená varianta matematického vztahu, jímž je zákon formulován, viz např. (Altmann, 1980), je

$$y = A.x^{-b}, \tag{15}$$

kde *x* je délka konstruktu naměřená v jeho konstituentech, *y* je průměrná délka konstituentu v jednotkách na nejbližší nižší jazykové rovině a *A, b* kladné parametry. Úplná verze matematické formule MAL, jejíž testování bude také dále popsáno, je

$$y = A.x^{-b}.e^{cx},^{3} \tag{16}$$

kde *A, b* jsou kladné parametry a *c* parametr záporný, viz (Altmann, 1980) a (Altmann et al., 1989).

Zásadním přínosem Luďka Hřebíčka bylo, že za prvé MAL platí na různých lingvistických úrovních stejně, což nazval soběpodobností[4]. Za druhé díky platnosti MAl na různých hladinách prokázal existenci nadvětných struktur. A za třetí si povšiml zásadní souvislosti Moranovy formule, vzorce pro výpočet fraktální dimenze a MAL. Pro další detaily viz např. (Hřebíček, 1997) a (Hřebíček, 2002).

Považuji za nutné nejprve zmínit samotné odvození Menzerath-Altmannova zákona a poté detailněji popsat fungování Menzerath-Altmannova zákona v našem experimentu.

Připomeňme si nejprve znění Menzerath-Altmannova zákona: čím delší je konstrukt, tím kratší jsou jeho konstituenty. Tuto definici je možné interpretovat do jazyka matematiky následujícím způsobem: Předpokládejme, že *x* je délka konstruktu a *y* délka konstituentu, pak relativní změna délky složek (tedy konstituentů) $\frac{\mathrm{d}y}{y}$ je dle zmíněného zákona úměrná relativní změně délky konstruktu $\frac{\mathrm{d}x}{x}$, tedy platí

$$\frac{\mathrm{d}y}{y} \sim \frac{\mathrm{d}x}{x}\ , \tag{17}$$

tedy pokud stanovíme $b > 0$ jako koeficient úměry, pak

$$\frac{\mathrm{d}y}{y} = -b.\frac{\mathrm{d}x}{x}\,. \tag{18}$$

(18) je diferenciální rovnice, kterou snadno vyřešíme metodou separace proměnných. Obecným řešením je rovnice

$$\ln|y| = -b.\ln|x| + c, \text{ kde } c \text{ je reálná konstanta.} \tag{19}$$

[3] Již dříve bylo ukázáno, že role exponenciální části, jež odlišuje zkrácenou a úplnou formuli MAL, narůstá s klesajícími lingvistickými úrovněmi, tzn., neměla by být vypouštěna při studiu slov a slabik, a naopak může být opomenuta u vyšších hladin, jako jsou věty, klauze, sémantické konstrukty, viz (Andres, 2010).

[4] Použití termínu soběpodobnost v tomto smyslu se signifikantně liší od jeho významu chápaného v souvislosti s fraktální geometrií. Na základě tohoto nesouladu by mohly být bez dalšího dodefinování mylně přiřazovány objektům fraktální vlastnosti. Pochybnosti budou rozptýleny dále v textu pomocí dalšího aparátu.

Jelikož *x* a *y* jsou nezáporná čísla, můžeme odstranit absolutní hodnoty a substituovat parametr $A = e^c$, dostaneme tak jednoduchou variantu formule Menzerath-Altmannova zákona (15)

$$y = A.x^{-b},$$

viz např. (Wimmer, 2003) a (Hřebíček, 1997).

Pro tento experiment byly zvoleny následující, zřetelně definované binarismy[5], které odpovídají jednotkám v experimentech Luďka Hřebíčka, viz (Hřebíček, 1997), (Hřebíček, 2002), (Andres, 2009) a (Wimmer et al., 2003), a především pojetí stanovení jednotek je pojednáno dále detailněji a je navrženo několik možných přístupů, které jsou diskutovány. Binarismy v našem experimentu jsou následující: sémantický konstrukt[6] (měřený v počtu vět) – věty/klauze/syntaktické konstrukce (měřené v průměrné délce svých slov)[7], věty/klauze/syntaktické konstrukce (měřené v počtu slov) – slova (měřená v průměrné délce jejich slabik[8]) a slova (měřená v počtu slabik) – slabiky (měřené v průměrné délce jejich fonémů)[9]. Všechny výše zmíněné jednotky potřebují detailní zadefinování, které bude poskytnuto dále v textu.

Přeložme si nyní zmíněné vztahy do jazyka matematiky. Nechť *i* je přirozené číslo, pro náš experiment předpokládáme $i = 1,2,3$, což představuje tři námi zavedené lingvistické binarismy: $i = 1$ pro vztah sémantický konstrukt – věta, $i = 2$ pro vztah věta – slovo a $i = 3$ pro vztah slovo – slabika. Dvě výše zmíněné varianty Menzerath-Altmannova zákona mohou být tedy přesně zapsány jako

$$y_i = A_i.x_i^{-b}, \text{ pro každé } i = 1,2,3, \qquad (20)$$

nebo rozšířená verze MAL

[5] Binarismy se v kontextu tohoto experimentu rozumí vztahy mezi dvěma bezprostředně ležícími jazykovými hladinami.

[6] *Sémantický konstrukt* je Hřebíčkem nově navržená nadvětná jazyková struktura. Luděk Hřebíček sám pro ni navrhoval název agregát. Proti tomuto návrhu se objevila kritika. Gabriel Altmann navrhl posléze termín hreb, viz (Hřebíček, 1997) a (Hřebíček, 2002). Prozatím ale pro tuto mladou nadvětnou strukturu ponechme označení sémantický konstrukt.

[7] Mezi první a druhý zmíněný binarismus je navrženo v dalších zkoumáních vložit vztah: věta (měřená v klauzích) – klauze (měřená v průměrné délce slov v nich). Posloupnost binarismů bude po této úpravě následující: sémantický konstrukt (ve větách) – věta (v průměrné délce klauzí), věta (v klauzích) – klauze (v průměrné délce slov), klauze (ve slovech) – slovo (v průměrné délce slabik), slovo (ve slabikách) – slabiky (v průměrné délce fonémů). Viz například (Buk & Rovenchak, 2008).

[8] Může vzniknout pochybnost, zda na tomto místě použít slabiky či morfy. Jelikož se snažíme postihnout míru sémantičnosti v textu, jevilo by se jako smysluplnější použít morf spíše než slabiku, jelikož má jasnou sémantickou funkci, viz (Hřebíček, 2002). V závěrečném zhodnocení své práce navrhuji další způsoby, jaké jednotky vzít v úvahu a z jakého důvodu. Prozatím se pro jednoduchost omezím na výše zavedenou posloupnost binarismů jazykových rovin a podržím se původního Menzerathova zkoumání. Obdobná diskuse by mohla proběhnout při úvaze, zda použít hlásky, fonémy či grafémy. O volbě jednotek bude pojednáno dále v textu.

[9] Jelikož nám jako nástroj ke zkoumání textů slouží fraktální teorie, což je, jak bude ukázáno dále, teoretická metoda založená na nekonečně mnoha aproximacích, je vhodné v dalších budoucích experimentech pokud možno rozšířit počet zkoumaných jazykových úrovní směrem nahoru i dolu. V závěru budou navrženy možnosti dalších experimentů.

$$y_i = A_i . x_i^{-b} . e^{c_i x_i}, \text{ opět pro každé } i = 1,2,3. \quad (21)$$

Cílem našeho experimentu je vykonat fraktální analýzu textu výše a dále naznačeným způsobem. Jde tedy o testování hypotézy a provádění experimentů. Na tomto základě je též budována metodologická základna. Jako objekty experimentu byly zvoleny jeden poetický text v originále i v překladech, viz (Poe, 1985), (Poe, 1993), (Poe, 2008), a jeden žurnalistický text, (Nebeský, 2009). Texty budou analyzovány pomocí Menzerath-Altmannova zákona, kde pro naše účely bude signifikantní určení parametru b_i, pro každé $i = 1,2,3$.

Zde opět nastává chvíle, kdy je třeba se zmínit o historickém vývoji a zároveň souvislosti mezi Menzerath-Altmannovým zákonem a fraktální teorií. Jak již bylo zmíněno, velkou zásluhou Luďka Hřebíčka byla prokázána jednoduchá souvislost mezi MAL a fraktální dimenzí, která je podstatou definice fraktálu č. 1.

Pokud si všimneme korelace mezi veličinami figurujícími ve vztahu pro výpočet fraktální dimenze a veličinami vystupujícími v MAL a substituujeme v rovnici (13) $N = x$ a $r = y$, dostáváme

$$D = \frac{\ln x}{\ln \frac{1}{y}} = \frac{\ln x}{-\ln y}, \quad (22)$$

tedy

$$\ln x = -D . \ln y, \quad (23)$$

$$\ln y = -\frac{1}{D} . \ln x . \quad (24)$$

Křivka, která je geometrickou interpretací této rovnice, může být beze změny svého sklonu libovolně vzdálena od osy *x*. Tato vlastnost budiž vzata v úvahu pomocí následující korekce rovnice, zároveň zaveďme substituci $b = \frac{1}{D}$ (24):

$$\ln y = -b . \ln x + \ln A . \quad (25)$$

Tuto rovnici již velice snadno převedeme pomocí vět o logaritmech na tvar Menzerath-Altmannova zákona

$y = A . x^{-b}$, viz (Hřebíček, 1997) a (Hřebíček, 2002).

Z výše uvedeného nyní již zcela jasně vyplývá souvislost parametru *b* Menzerath-Altmannova zákona a fraktální dimenze *D.* Aby se ale dalo uvažovat o spojitosti s fraktální dimenzí a fraktální teorií, je nutné, aby byly splněny dvě podmínky. První z nich je, že $b_i > 0$ pro všechna $i = 1,2,3$. A dále reciproká hodnota aritmetického průměru všech parametrů b_i, $i = 1,2,3$,

$$D = \frac{3}{b_1 + b_2 + b_3} \quad (26)$$

může být interpretována jako soběpodobnostní dimenze přidruženého matematického fraktálu[10], který může být aproximován s dostatečnou přesností vizualizovaným modelem jazykové struktury, která je analyzována.

Následně tedy *jazykový fraktál* může být definován jako takový lingvistický subjekt, který splňuje Menzerath-Altmannův zákon se všemi b_i na všech zkoumaných jazykových úrovních pozitivními. V porovnání s v principu lineárním (tj. jednodimenzionálním) de Saussureovým pojetím výpovědi či textu, číslo *D*, soběpodobnostní dimenze textu, tudíž odráží *míru sémantičnosti* textu, viz (Andres, 2009).

Je třeba též zdůraznit, že není možné splnit očekávání, že bude dokázáno, že jazyková struktura je matematickým fraktálem, protože počet zkoumaných jazykových úrovní je a bude konečné číslo, ať se budeme snažit jakkoli, viz (Andres, 2010), (Köhler, 1995) a (Köhler, 1997). Tudíž, pravděpodobnost jazykové fraktality je pro nás výzvou v aproximativním a statistickém slova smyslu. Přesto, jak již bylo zmíněno dříve, nevylučujeme, ale naopak usilujeme o potenciální rozšíření počtu zkoumaných jazykových úrovní. Zásadní pojetí spočívá ve faktu, že všechny modely jsou cyklicky soběpodobné. Jedním cyklem rozumíme tři zkoumané jazykové úrovně; tj. tři jazykové úrovně v našem experimentu se rovnají jedné iteraci při konstrukci fraktálu. Postup, který jsme zvolili v našem experimentu, je následující: Poté, co dále zmíněnými způsoby najdeme parametry A_i, c_i a obzvláště b_i, zjistíme, zda je daný výběrový soubor jazykovým fraktálem dle výše zmíněných kriterií. Pokud ano, spočteme dimenzi D[11] přidruženého matematického fraktálu a při jeho konstrukci[12], vizualizaci, vyjdeme z definice 2. Jazykový fraktál je tedy jistou aproximací tohoto fraktálu matematického.

3.1.3 Další ohlasy na teorii fraktálních struktur v jazyce

Poté, co jsme se blíže podívali na vznik teorie fraktálů v matematice, Altmannovu a Hřebíčkovu teorii fraktální povahy jazyka, bych ráda zmínila několik dalších aplikací této teorie v lingvistice.

Na úvod se opět chci vrátit k *Benoitu Mandelbrotovi*, který ve svém díle reagoval a upravil výše již zmiňovaný **Zipfův zákon** týkající se frekvencí slov (Mandelbrot, 2003). Úprava se proto často nazývá **Zipf-Mandelbrotův zákon**. Mandelbrot sestavil takzvané **lexikografické stromy**, jejichž struktura není sice naprosto totožná jako struktura jazyka, ale jedná se opět o matematicky dokonalou konstrukci, která se obvykle ve skutečnosti nevyskytuje.

Slovní zásoba je chápána jako množina posloupností písmen, které jsou akceptovatelné jako slova. Tato slova jsou od sebe oddělena mezerami. Lexikografické stromy jsou sestrojeny následujícím způsobem: Kmen stromu reprezentuje mezeru, dělí se pak dále na *N* větví první úrovně, které odpovídají každému písmenu dané abecedy. Každá tato větev se dělí na *N* větví

[10] Pomíjíme tedy definici 1. fraktálu, neboť připouštíme i celočíselnou dimenzi matematického fraktálu.
[11] Luděk Hřebíček ve své původní teorii zjišťoval dimenzi na každé jednotlivé jazykové hladině. Aby v takém případě byl daný výběrový soubor prohlášen fraktálem, musely by dimenze na všech jazykových hladinách být přibližně stejné, což je případ extrémní a vyjímečný, jak bude vidět z výstupů našeho experimentu. Jan Andres ve své formalizaci Hřebíčkovy teorie definoval dimenzi *D* výše zmíněným způsobem, což umožňuje modelovat přidružený matematický fraktál se stejnou dimenzí. Typ matematického fraktálního objektu byl zvolen Cantorovský.
[12] Obvyklým postupem je najít dimenzi známého fraktálního objektu.

druhé úrovně atd. Každé takové rozvětvení stromu reprezentuje slovo, které může být opatřeno pravděpodobností existence takového slova v jazyce. Z teorie pravděpodobnosti a původního Zipfova zákona dostáváme:

$$1+N+N^2+\ldots+N^{k-1}=\frac{N^k-1}{N-1}<\rho<\frac{N^{k+1}}{N-1}$$

Zavedeme substituci:

$$D=\frac{\log N}{\log\frac{1}{r}}$$

$$V=\frac{1}{N-1}$$

$$k=\frac{\log\frac{P}{P_0}}{\log r}$$

Po dosazení tedy dostaneme:

$$(P^{-D}\cdot P_0^D)-1<\frac{\rho}{V}\leq N\cdot(P^{-D}\cdot P_0^D)-1$$

A tedy:

$$P=P_0\cdot\left(\rho+V\right)^{-1/D}$$

N počet písmen abecedy

K počet úrovní konstrukce

ρ rank slova s pravděpodobností *P*

P_0 činitel zajišťující, že součet všech pravděpodobností je 1

D,V nezávislé parametry

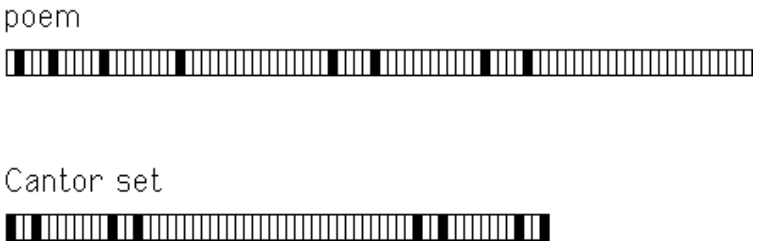
Pokud *D<1,* pak jde o fraktální dimenzi. Jestliže $D\geq 1$, pak je ρ omezené a tzn. slovní zásoba obsahuje konečný počet slov. *D* je také mírou bohatství slovníku.

Geometrická interpretace je taková, že máme na intervalu $\langle 0,1\rangle$ *N* intervalů délky spojené s *N* písmeny abecedy. Každý interval („písmeno“) se dělí na *N* intervalů („písmeno-písmeno“) a jeden interval („písmeno-mezera“) atd. Interval „mezera“ se dále nedělí a definuje posloupnost písmen končící mezerou. Jestliže ztotožníme mezeru s dírou, pak doplněk takto definovaných děr je Cantorovou množinou s dimenzí D.

Jak uvádí *Lynellen D.S. Perry* každá větev lexikografického stromu je v redukovaném měřítku celý strom. Avšak běžné mluvené i psané jazyky nevyrůstají na takovýchto stromech,

a pokud na tom přesto budeme trvat, pak je většina jejich větví mrtvých, viz [2]. K fraktální struktuře přirozených jazyků se dále také hlásí *Benny Shanon* z Hebrew University v Jeruzalémě ve své práci „Fractal Patterns in Language“.

Zajímavá je též práce *Lucy Pollard-Gott* „Fractal Repetition Structure in the Poetry of Wallace Stevens“, ve které nachází podobnost poezie s Cantorovou množinou. Postup je následující. Ze Stevensových básní vybrala klíčové slovo. Každé slovo pak na ose nahradila rámečkem, přičemž začernila rámeček v pořadí klíčového slova. Uveďme si zde příklad platný pro báseň The Sail of Ulysses, viz obr. č. 5.



Obr. č. 5: Porovnání struktury básně The Sail of Ulysses a Cantorovy množiny

Co by pro Stevensovu poezii fraktální struktura měla znamenat? Hierarchie je důležitý aspekt jazyka i hudby. Stevensova poezie je považována za velice muzikální, což je jistě způsobeno také hierarchií opakování. Navíc je jistě na místě připomenout Hřebíčkovu teorii, že opakování klíčového slova činí agregáty kompaktními. Dále pak zajisté opakování uvádí v mysli v pohyb nekonečný ústup ke stále nižším škálám, aby tak posílilo na pár řádcích „vstup do jiného světa“, viy [3].

Dalším, koho bych zde chtěla jmenovat, je *Edda Leopold*, viz [4]. Jeho práce je přímou reakcí na Hřebíčkovu teorii a přináší rozbor a matematický aparát pro tuto teorii. Hřebíček se v této části své teorie snaží exaktně potvrdit existenci dalších úrovní, do té doby filology neuvažovaných. O konstituentech a konstruktech jsem se již zmínila. O jejich vztahu dále platí, že každá jazyková úroveň je zároveň vůči vyšším jazykovým úrovním konstituentem a vůči nižším konstruktem. Mezi jazykovými úrovněmi existuje tudíž podobnost vyjádřená Menzerath-Altmannovým zákonem. Konstituent závisí na konstruktu, konstrukt na konstituentu. Proto můžeme Menzerath-Altmannův zákon upravit do podoby iterativního vyjádření řetězce jazykových úrovní:

$$x_1 = \left(\frac{A_1}{\left(\frac{A_2}{\left(\frac{A_3}{x_4} \right)^{1/b_3}} \right)^{1/b_2}} \right)^{1/b_1} \tag{27}$$

x_1 nejvyšší úroveň jazyka, x_4 nejnižší úroveň

Výhodou tohoto vztahu je, že je rovnicí pouze jedné proměnné. Vztah vyjadřuje různé strukturní části představované úrovněmi. Hřebíček uvažuje dále text jako časovou řadu a analyzuje ji Hurstovými indikátory. (Hřebíček, 2002)

Leopold uvažuje fraktál jako podmnožinu Eukleidovského prostoru – vnořený prostor (angl. imbedding space). Ukazuje ve své práci, že fraktální interpretace Menzerath-Altmannova zákona vede ke zcela abstraktnímu vnořenému prostoru, který nemá metriku. Upravuje dále pro potřebu analýzy textu ve smyslu Hřebíčkovy teorie vzorec pro výpočet Hausdorffovy dimenze, který také s pomocí Falconera (Falconer, 1990) komentuje.

Mějme množinu *B*, která je podmnožinou Eukleidovského prostoru R^n, pro každé $\delta > 0, s \geq 0$

$$H_\rho^s(B) = \inf\left\{\sum_{i=1}^{\infty} |U_i|^s : \{U_i\} \text{ je } \rho \text{ uzávěr } B\right\}$$

Hausdorffova dimenze *B* je kritická hodnota *D*, kde $H_s(B)$ osciluje mezi ∞ a 0. Platí

$$H_s(B) = \lim_{\delta \to 0} H_\delta^s(B) = \begin{cases} \infty & jestliže \quad s < D \\ 0 & jestliže \quad s > D \end{cases}.$$

Podle Falconera, je-li *B* konečná nebo spočetná množina, pak je Hausdorffova dimenze rovna nule, což znamená, že pozorovaná data nikdy nemohou mít Hausdorffovu dimenzi různou od nuly, protože nikdy nejsme schopni uskutečnit nekonečný počet pozorování. Když tudíž řekneme, že data reprezentují fraktální strukturu, je to vždy idealizace v tom smyslu, že pozorovaná struktura je přibližně dotvářena do nekonečně malé míry.

Hřebíček také upozorňuje na zcela zjevnou soběpodobnost výše zmíněného iterativního vyjádření řetězce jazykových úrovní a přirovnává vztahy mezi různými analyzovanými subsystémy jazyka v Menzerath-Altmannově zákoně ke generátoru Cantorovy množiny. Existence fraktální struktury se tudíž zdá zjevným důsledkem Menzerath-Altmannova zákona, ale je obtížné tuto hypotézu uchopit exaktně, jak uvádí Leopold. Dále vysvětluje, že abychom definovali fraktální dimenzi z Menzerath-Altmannova zákona, je potřeba spojitá škála úrovní k analýze. Tudíž bychom měli být schopni pokračovat nepřetržitě od subsystému fonémů k morfémům (či slabikám), dále ke slovům, klauzím, větám a nadvětným strukturám. Dále jestliže δ označuje úrovně analýzy na spojitém svazu, pak musí být definována limitní úroveň analýzy pro $\delta \to 0$. Pokud je to splněno, pak může být definice Hausdorffovy dimenze přizpůsobena Hřebíčkovým idejím fraktálních struktur v textu.

V neposlední řadě je nutné připomenout *Aliho Eftekhariho* z Electrochemical Research Center v Teheránu a jeho esej „Fractal Geometry of Literature: First Attempt to Shakespeare's Works", viz (Eftekhari, 2006). Eftekhari se v této práci zabývá fraktální analýzou písmen. Odkazuje na práce *K. J. Hsu* a *A. Hsu*, kteří se zabývají prokazováním fraktálního charakteru hudby a výpočtem její dimenze. Na druhé straně je také možné postupovat opačným směrem a stvořit hudbu podle fraktálního principu, což se zatím zdá vůči literatuře utopické, neboť tento postup postrádá sémantickou složku. Nicméně na základě podobnosti textu a hudby

(linearita zápisu) je zkoumána literatura. Pro výpočet fraktální dimenze textu je použita stejná formule jako pro hudbu:

$$F = \frac{c}{i^D} \tag{28}$$

D	fraktální dimenze skladby	fraktální dimenze literatury
i	interval mezi dvěma následujícími tóny	interval mezi dvěma písmeny v abecedních řadách
F	procento frekvence i	procento frekvence i
c	konstantní proporční faktor	konstantní proporční faktor

V abecedě si před písmenem A představíme vymyšlené prázdné funkční písmeno, pro něž *i=0*, dále pak $i_A = 1, i_B = 2, \ldots, i_Z = 26$. *i* je tedy totožné s rankem v abecedních řadách. Aplikací předchozího na díla Williama Shakespeara zjistil jednak, že výskyt jednotlivých písmen v textech má chaotický charakter, a jednak spočetl fraktální dimenze jednotlivých textů. Jejich velikost se pohybuje mezi 0,4500 (Hamlet) a 0,5985 (Macbeth). Připomeňme, že fraktální dimenze Cantorovy množiny je 0,6309. Rozdílná dimenze může sloužit podle Eftekhariho například k porovnávání jednotlivých děl.

Další důležitý pojem, který Efekhari zmiňuje, je faktor fraktality $\zeta \in \langle 0,1 \rangle$, který určuje, jak moc je povrch objektu definován fraktálními strukturami. Pro reálné objekty je faktor nižší, klesá například korozí nebo faktorem hrubosti objektu. Fraktálně generovaná hudba má $\zeta = 1$, ale text, jak jsem již zmínila, prozatím fraktálně vygenerován nebyl.

4 Algoritmus kvantitativní analýzy textu

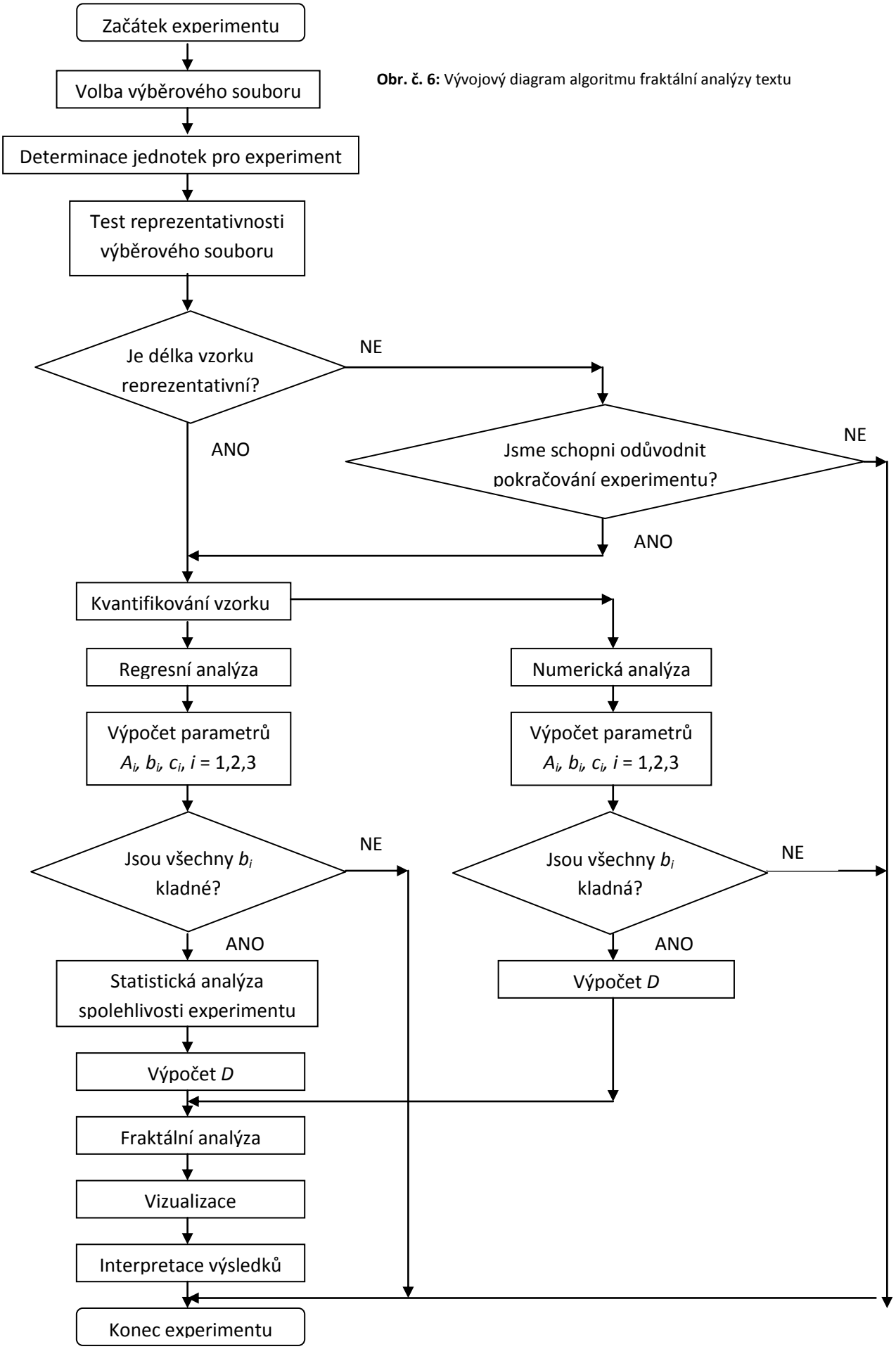
Hlavním důvodem pro tento experiment je vyslovit teorii jazykových fraktálů a podpořit ji dostatečným množstvím experimentů. Dále je možné ukázat, že míra sémantičnosti zkoumaných textů může být definována a měřena prostřednictvím fraktální dimenze. Dalším důvodem je prezentovat způsob, jakým vizualizovat textový výběrový soubor prostřednictvím MAL a dalších nástrojů fraktální teorie. V neposlední řadě vyvstává jako velice důležité sestavit algoritmus, který by posloužil pro kvantitativní zpracování dalších textů a pro následné vyhodnocení získaných výsledků autorce i dalším nadšencům z řad lingvistů i matematiků. Následující část této práce je rozdělena na sekce, které čtenáře povedou logicky a detailně jednotlivými kroky algoritmu. V této souvislosti je též připraven vývojový diagram tohoto algoritmu. A na závěr bych ráda poukázala na některé problémy, které mohou v různých úrovních tohoto experimentu nastat a naznačila některé závěry, které z celé práce plynou, a pokusím se poukázat na jednotlivé problémy. Tato kapitola je rozdělena na část teoretickou, která je nutně doplněna praktickou sekcí s ukázkami výpočtů v každém svém jednotlivém kroku a komentářem, přehledy všech výstupů ve formě tabulek, grafů a obrázků jsou zařazeny do příloh.

4.1 Teoretické poznámky k algoritmu kvantitativní analýzy textu

Celá procedura vyšetřování textu kvantitativním způsobem se skládá z následujících nutných kroků, algoritmus je diskutován též v (Andres, 2009), (Andres et al., 2011) a (Andres & Benešová, 2011).

1. Nejprve je nutné pečlivě zvolit text, který bude podroben analýze.
2. Pečlivě stanovíme jednotky, se kterými budeme dále operovat.
3. Ověříme reprezentativnost výběrového souboru. Při určitých odhadech parametrů základního souboru je důležité, aby výběr byl *reprezentativní*.
4. Kvantifikujeme text, abychom z něj extrahovali proměnné x_i a y_i pro každé $i = 1,2,3$, k čemuž použijeme klasifikované a pevně stanovené jazykové jednotky, viz bod 2.
5. Na základě dále v textu detailněji popsaných statistických metod (regresní analýza) a numerických metod najdeme parametry MAL A_i, c_i, obzvláště b_i pro každé $i = 1,2,3$ a reciprokou hodnotu jejich aritmetického průměru *D*.
6. Musí být opět pomocí statistických metod otestována spolehlivost modelu celého experimentu.
7. Parametry musí být interpretovány ve fraktální analýze.
8. Provedeme vizualizace jazykových struktur pomocí postupných aproximací matematických fraktálů s danou dimenzí *D*, popřípadě též shlukovou analýzu.
9. Vizualizace jazykových struktur, výsledky experimentu i shluková analýza musí být interpretovány.

Výše popsaný algoritmus je shrnut a vizualizován ve vývojovém diagramu na obr. č. 6. Jednotlivé kroky algoritmu jsou kromě svého detailního popisu doplněny v praktické části této kapitoly výpočty aplikovanými na zvolené výběrové soubory.



Obr. č. 6: Vývojový diagram algoritmu fraktální analýzy textu

4.1.1 Krok 1 – volba výběrového souboru

V ideálním případě bychom usilovali o analýzu tzv. základního souboru/populace, což je množina objektů, které chceme systematicky a obecně popsat. Obvykle však k dispozici nemáme všechny prvky této množiny a musíme se omezit na tzv. výběrový soubor/vzorek, který je podmnožinou základního souboru a měl by co nejlépe vystihovat jeho vlastnosti, které chceme popsat. Odhadujeme tedy realitu základního souboru co nejpřesněji pomocí poznání nějaké jeho části, viz (Volín, 2007).

Někdy se ukazuje jako velice složité stanovit, co je základní a co je výběrový soubor, dokonce dle Orlova, viz (Orlov et al., 1982), základní soubory neexistují. Dle (Wimmer et al., 2003) základní soubory existují, ale některé z nich jsou nespočetnými množinami. Pokud ale existují i menší základní soubory, je nutné je pečlivě nadefinovat a uvést, co přesně představují.

Dle Marie Těšitelové, (Těšitelová, 1987), je při výběru materiálu pro zkoumání nutno brát v úvahu kritéria kvalitativní – respektující kriteria jazyková, psychologická, sociologická, tematická, „sémiotická“ a jiná – a kvantitativní, což je způsob, jakým provádíme výběr, a výsledek činnosti vybírání, viz krok 3.

4.1.2 Krok 2 – stanovení jednotek

Pro spolehlivý experiment, který má smysl verifikovat, je důležité pečlivě stanovit jednotky, které budou používány. Při stanovování jednotky je nutné respektovat čtyři základní pravidla:

1. Jednotka musí být jednoznačně definována, pokud jde o zvolené znaky, viz (Těšitelová, 1987, s.19).
2. Pojetí jednotky má být ve shodě s běžným pojetím v lingvistice, popřípadě maximálně přijatelným pojetím, viz (Těšitelová, 1987, s.19).
3. Vymezení jednotky souboru během práce důsledně zachováváme, aby byla zaručena maximální homogennost analyzovaného souboru a aby získané výsledky byly maximálně srovnatelné s analogickými pracemi, viz (Těšitelová, 1987, s.19).
4. Každá jednotka se někde započítává a žádná jednotka se nepočítá dvakrát, viz (Těšitelová, 1987, s.12).

V tabulkách uvedených níže je seznam délek konstruktů a konstituentů pro tři stanovené binarismy, které jsou podrobeny experimentu:

1. úroveň $i = 1$: x_1 sémantický konstrukt (jeho délka ve větách/klauzích), z_1 jejich četnost - y_1 věty/klauze (jejich průměrná délka ve slovech)

2. úroveň $i = 2$: x_2 věty/klauze (jejich délka ve slovech), z_2 jejich četnost - y_2 slova (jejich průměrná délka ve slabikách)

3. úroveň $i = 3$: x_3 slova (jejich délka ve slabikách), z_3 jejich četnost - y_3 slabiky (jejich průměrná délka ve fonémech)

Naneštěstí není proces stanovování jednotek jednoduchý a jednoznačný. V této práci bych chtěla demonstrovat především tři přístupy. Ve třech speciálních případech jsou výběrové soubory nahlíženy pomocí čtvrtého způsobu. Samozřejmě můžeme použít také další alternativní definice jednotlivých jednotek. Nicméně jakmile jednou použijeme konkrétní definici, musíme ji striktně dodržet po celou dobu fraktální analýzy, jak říká třetí zásada výše. Determinace jednotek je též popsána v (Andres et al., 2011) a (Andres & Benešová, 2011).

Jak bylo zmíněno výše, před zahájením experimentu je třeba precizně zadefinovat jednotky na všech jazykových úrovních, ve všech binarismech a jejich definici důsledně dodržovat po celou dobu. Jako pro tento experiment nejobtížnějšími jednotkami se ukázala slova. V následujících odstavcích budou všechny jednotky stanoveny a speciálně v případě slov budou uvedeny všechny čtyři výše zmíněné přístupy. Záměrem bylo držet se alespoň zpočátku jednotek uvažovaných původně Menzerathem při vyslovení MAL a Hřebíčkem v (Hřebíček, 1997). Další způsoby stanovení jednotek budou dále definovány.

Foném. Pro první zmíněný binarismus potřebujeme definovat slova, slabiky a fonémy. *Foném* je základní jednotka fonologické jazykové úrovně. Akustickým nástrojům přirozených jazyků je přiřazen význam, proto mají platnost znaků. Jazyky plní svou funkci, protože znakové nástroje se znakovou platností mají komplexní povahu. Jsou složeny z jednotek, které samy o sobě nejsou znaky. *Foném je tedy souhrnem fonických prvků, který umožňuje uživateli rozlišovat jednotlivé znaky*, viz (Petr et al., 1986a) a (Štekauer, 2000).

Slabika. Aby bylo možné vykonat akustickou analýzu, která závisí speciálně u jednotek na vyšší úrovni podstatnou měrou na jazykové analýze, jsme schopni rozlišit akustické jednotky na rozličných úrovních. Řeč se skládá z vět, které jsou nejmenší řečové jednotky konzistentní s ohledem na svůj význam. *Slabika je nejmenší jazyková jednotka, u které je vztah jejích komponent tak úzký, že segmentujeme-li proud řeči, nejsme schopni jej rozdělit na kratší úseky, které by mohly usnadnit pochopení řeči.* Navzdory faktu, že uživatelé jazyka jsou obvykle schopni segmentovat svou řeč a slova na slabiky, lingvistika dosud nebyla schopna jednoznačně se shodnout na přesné definici podstaty slabiky, viz (Petr et al., 1986a).

Slovo. Je dáno již tradicí, že základní jednotkou morfologie je *slovo*. Termín slovo má ale rozličné významy, pokud se na něj soustředíme z pohledu rozličných jazykových úrovní. V naší analýze pohlížíme na slovo ze dvou různých úhlů pohledu. Za prvé jej chápeme jako konstrukt, jehož konstituenty jsou slabiky v binarismu *x* slova – *y* slabiky, a za druhé jej vidíme jako konstituent, jehož nadřazeným konstruktem je věta/klauze v binarismu *x* věta/klauze – *y* slovo a obdobně v binarismu *x* sémantický konstrukt – *y* věta/klauze měřená v průměrné délce slov, které se v nich vyskytují. První pohled je pohled fonologický, kdy na slovo nahlížíme jako na *fúzi fonémů*, druhý je pohledem syntaktickým. I když chápeme slovo jako morfematickou a morfologickou jednotku, je nutné, abychom rozlišovali mezi pojetím slova jako skutečně vyčlenitelné jednotky textu, jako *série morfů*, nebo pojetím slova jako jednotky jazykového systému, kde systémové slovo – *lexém* – reprezentuje *celou množinu svých „textových slov"* – slovoforem. To je důležité pro všechny flektivní jazyky, český jazyk není výjimkou. Nazývejme tedy slovo chápané prvním způsobem *slovoforma* a slovo chápané druhým způsobem *lexém*. Tato problematika je detailně pojednána například v (Petr et al., 1986b).

Po vyhodnocení prvních získaných výsledků bylo zjištěno, že kromě klasických definic slova je pro náš experiment důležité přidat několik dalších požadavků nezbytných pro zpracování textu pomocí Menzerath-Altmannova zákona. **Přístup 0.**[13] k definici slova je simplifikovat jej na jednotku textu existující „mezi dvěma mezerami“, sled grafémů „mezi dvěma mezerami“. Takže slovní tvar ve striktním slova smyslu, *syntetická slovoforma*, je lineární segment v proudu řeči nebo textu charakterizovaný svou sémanticko-funkční, zvukovou a grafickou úplností. Je to nezávislá volná forma, což se projevuje její přemístitelností (samozřejmě je tento fakt omezen syntaktickými pravidly a jazykovou typologií), více viz např. (Petr et al., 1986b), (Andres et al., 2011) a (Andres & Benešová, 2011). Výstup této metody je demonstrován v tab. Č. 19_1, 19_2, 19_3, 20_1, 20_2, 20_3 v příloze I. Tento způsob analýzy je bezpochyby jednodušší pro shromažďování dat, ale nereflektuje analytické vlastnosti jazyků, v našem případě nejen anglického, ale ve velké míře i českého jazyka, nebere v úvahu ani vztahy mezi různými slovy definovanými tímto způsobem, z čehož v neposlední míře vyplývá, že se tento způsob příliš nehodí pro kvantifikování míry sémantičnosti, o které bude více řeč později. Výhodou tohoto způsobu tedy je, že uvedená definice vykazuje velkou jasnost při kvantifikaci, a nevýhoda, že s sebou přináší řadu problémů podmíněných jednak typologickým charakterem jazyka, jednak vztahy gramatickými a sémantickými, které se v něm uplatňují, viz také (Těšitelová, 1987).

Podle **přístupu I.** chápeme pojem slova jako *složené (analytické) slovoformy*. Může být, jinými slovy, definováno jako specifické spojení syntetických slovních tvarů, které funguje jako komplexní tvar plnovýznamového slova. Pouze jedna z jeho komponent je nositelem lexikálního významu, na druhé straně další komponenta/ostatní komponenty je nositel/jsou nositeli významu gramatického, viz např. (Petr et al., 1986b).

Jedním z konkrétních složitých problémů morfologie českého slovesa, které bylo potřeba vyřešit při aplikaci druhého přístupu na texty v českém jazyce, bylo rozlišení tvarů pasíva složeného typu od verbálního adjektiva se sponovým slovesem, s čímž jsme se v překladech básně *Raven* hojně setkávali a na což poukazovala také Těšitelová v (Těšitelová, 1987). Příkladem budiž „jsem přikován“ z Bejblíkova překladu, kdy bylo nutné posoudit, zda se jedná o tvar pasíva odvozený od slovesa „přikovat“ nebo o tvar identický s „být přikovaný“. Analogicky je třeba rozlišit tvar pasívního reflexiva od nezvratného slovesa, opět např. viz (Těšitelová, 1987).

Výstupy této metody jsou publikovány v tab. č. 21_1, 21_2, 21_3, …, 40_1, 40_2, 40_3 v příloze I. Originální anglický text Poeova *The Raven* i všechny jeho překlady byly kvantifikovány podle přístupu I.

Přístup III. je parciálně sémanticky a parciálně účelově motivovaný. Základní pravidlo o stanovení slova jako analytického slovního tvaru je přejato ze přístupu I., který je pak obohacen o další požadavky. Slova mající funkci gramatických modifikátorů jiných slov bez

[13] Toto číslování je zvoleno vzhledem k faktu, že zmíněný přístup byl zvolen pro zahájení experimentu, nenáročnou ilustraci metodiky kvantifikace výběrových souborů, ale dále se neukázal být efektivní a vzhledem k volbě jednotek zcela lingvisticky korektní. Čili toto pojetí porušuje pravidlo druhé při stanovení jednotek.

ohledu na jejich ortografii jsou počítány jako části odpovídajících slovních tvarů, viz např. (Hřebíček, 1997). Z definice pádu vyplynulo, že není možné instalovat v gramatickém popisu předložku jako slovo; ta proto ztrácí slovnědruhovou příslušnost a stává se výrazověobsahovou součástí pádového systému, viz (Faltýnek, 2011). Tudíž předložky modifikující řídící substantivum jsou počítány jako jedna jednotka dohromady s bezprostředně následujícím slovem, ať už je to řídící jméno nominální vazby či ne. Důvodem pro výběr bezprostředně následujícího slova je, že výběr korektní předložky je determinován výchozím fonémem právě následujícího slova kvůli výslovnosti (například v české *v čem* x *ve vesnici*), viz (Andres et al., 2011).

V případě, že bychom počítali předložky jako samostatné slovní tvary, v binarismu slova (v počtu svých slabik) – slabiky (měřené v průměrné délce fonémů) by neslabičné předložky v českém (a některých dalších slovanských) jazyce musely být počítány s $x = 0$[14]. Dle (Wimmer et al., 2003) je doporučeno neslabičné předložky z výpočtů vynechat nebo je připočítat k délce dalšího slova, o čemž bylo již pojednáno výše. Vynechání předložek není zde doporučeno, neboť by tímto způsobem došlo ke ztrátě dat (fonémů při výpočtu délky slabiky).

Společným problémem jazyků s „chudou" morfologií, pokud jde o pojetí slova jako jednotky, je člen (určitý a neurčitý). Podle Těšitelové se pokládá zpravidla za samostatnou jednotku, viz (Těšitelová, 1987, s. 15). Tento přístup byl prozatím při rozboru originálního textu respektován. Nicméně, v rámci rozboru metodiky stanovování jednotek je navrženo, aby byl člen v dalších experimentech počítán jako jedna entita se svým řídícím substantivem či jeho substitutem, jak ostatně navrhuje též Hřebíček: „Words having the function of grammatical modifiers of words, regardless of their orthography, were counted as parts of the respective word forms. For example, indefinite article and postpositions – with the exception of those connected with the modified words by genitive construction – were counted as parts of the modified words and not as separate words," pro další detaily viz (Hřebíček, 1997, s. 18). Tato definice je podpořena též v (Dušková, 1994, s. 61): „Syntakticky má člen funkci determinátoru, tj. nesamostatného větného členu v rámci větného členu realizovaného substantivem, v němž zpravidla tvoří první složku, tj. předchází před premodifikací. Od premodifikátoru se dále liší tím, že příslušný větný člen provází obligatorně (může mít ovšem nulovou podobu) a že je pouze jeden, kdežto modifikátory jsou vždy fakultativní a může jich být několik."[15] [16]

[14] Tento fakt by mimo jiné způsobil problémy při výpočtu koeficientů b_i, jak bude patrné v následující kapitole. Neslabičné předložky se skládají z jednoho konsonantu, čili by tedy pro $x = 0$ musela průměrná hodnota jeho konstituentů být $y = 1$.

[15] Zde je nutné připojit několik poznámek o problémech, které mohou nastat, jestliže budeme spojovat členy některým ze členů fráze, kterou modifikují. Za prvé, člen bude záhodno počítat dohromady s bezprostředně následujícím slovem, a nikoli nutně s řídícím členem fráze, který není bezprostředně následující, z obdobných důvodů zmíněných výše v souvislosti s počítáním předložek. Tedy například v případě fráze „an unseen censer", jelikož neurčitý člen v angličtině má dvě varianty *a/an*, je nutné pojit dohromady „an unseen" a „censer" zvlášť. Kdybychom pojili dohromady člen s řídícím substantivem fráze, byla by volba varianty neurčitého členu *an* nekorektní.

[16] V úvaze o členech nesmíme zapomínat na to, že v angličtině existují členy tři: určitý, neurčitý a nulový. V případě, že bychom počítali se členy, pak je nutné vyřešit opět případ nulového členu, neboť v takovém případě opět v binarismu slovo ve slabikách – slabiky ve fonémech dostáváme $x = 0$, což již bylo diskutováno výše v textu.

Přístup II. jde ruku v ruce s přístup III. Jediný rozdíl je, že určitý a neurčitý člen v anglickém a německém jazyce byl chápán a definován jako součást jednotky „slovo“.

Věta. Většina psaných textů i promluv má komplexní povahu, tzn., můžeme je segmentovat na elementární textové jednotky, které je možno oddělit a identifikovat v promluvě akustickými signály a v psaném textu grafickými signály (tečka, otazník, vykřičník nebo dvojtečka). *Věta představuje komplexní strukturu v ohledu formálně gramatickém i sémantickém.* Organizačním centrem této struktury je predikát. To je jazyková jednotka, která je ve své větotvorné funkci realizována jako jakýkoli finitní slovesný tvar, výjimečně též jako infinitiv, viz (Petr et al., 1987).

Existuje velké množství definic věty. Pro účely našeho experimentu se omezím na dvě základní pojetí. První z nich je aplikováno v praxi v našem experimentu. *Věta je chápána jako predikační jednotka; vztahuje se k určitému slovesu*, viz (Těšitelová, 1987, s.16). Tento přístup nepřináší v českém jazyce větší problémy. Předmětem našeho zkoumání je ale poetický text, jehož syntaktická struktura je podstatně volnější než struktura ostatních stylů.

Problém však nastává při analýze textu v anglickém jazyce. Česká vedlejší věta má v angličtině tři ekvivalenty, vedlejší větu, gerundiální a infinitivní vazbu. Oproti hojnému používání vedlejší věty v češtině, v angličtině je její použití až na třetím místě, a to pouze v jistých, přesně určených situacích. Větné členy vyjádřené jmennými tvary mohou často též nabýt větné formy beze změny významu. Důvodem rozdílnosti v obou jazycích je systém formálních prostředků, který je v angličtině obohacen o gerundium a je více rozvinut, neboť jmenné tvary mohou vyjadřovat bohatý systém kategorií, jako je slovesný rod, čas, aktivum či pasivum atd. Větné členy vyjádřené jmennými tvary slovesnými obsahují sekundární predikaci, viz (Dušková, 1994, s. 542). Proto je zcela nezbytné vyřešit problém, zda počítat konstrukce řízené jmennými tvary slovesnými, které představují ekvivalent českým větám vedlejším, jako věty ve smyslu první definice. Přikláníme se k názoru, že je to vhodné, viz též Hřebíček v (Hřebíček, 1997, s. 18): „Sentences were taken as text segments having finite and infinite verbs as their heads; …. . On the other hand, gerunds (nebo infinitivy, pozn. autorky) standing in a sentence close by a finite verb of the same sentence were not classified as sentence heads.” Tab. č. 19, …, 46 v příloze I. přinášejí výsledky „českého pojetí“ zpracování originálního textu, kdy striktně vyžadujeme, aby řídící člen každé věty byl predikát realizovaný finitním tvarem slovesa. Druhý navržený způsob je ilustrován v tab. č. 47 v příloze I.[17]

Další možné pojetí věty je mechanické a do jisté míry odpovídá pojetí slova jako grafické jednotky a věta je chápána jako slovo nebo skupina slov „od tečky k tečce“ či od „velkého písmena na začátku věty k finálnímu interpunkčnímu znaménku“, viz (Těšitelová, 1987, s. 16). Toto pojetí doporučujeme zejména v případě, že bychom vkládali o jednu úroveň a tudíž zároveň o jeden binarismus více, věta (měřená v klauzích) – klauze (měřená v průměrné délce slov v nich), jak bylo zmíněno výše, viz poznámka 7.

Sémantický konstrukt. „Věty v textu, které obsahují jistou lexikální jednotku/lexém (a tvoří tak širší kontext jednotlivé lexikální jednotky) jsou jazykové konstrukty těchto

[17] Třetí úroveň vykazuje pro obě dvě pojetí stejné výsledky, neboť se v ní neobjevují věty ani jako konstrukty ani jako konstituenty.

odpovídajících konstituentů, tedy vět," viz (Hřebíček, 1997, s. 31). To je způsob, jakým Luděk Hřebíček zavedl jazykovou hladinu, která se nachází nad syntaktickými strukturami. Nazýval takový konstrukt agregátem, ale tento termín nebyl obecně přijat. Nazývejme takovouto jazykovou strukturu prozatím termínem *sémantický konstrukt*. Povaha sémantického konstruktu je mírně odlišná od povahy jednotek na nižších jazykových úrovních. Každá věta se sestává z *n* (položme $n \geq 1$ pro český, anglický i německý jazyk) lexémů. Tudíž každá věta náleží *n* sémantickým konstruktům jako jeden z jejich konstituentů (pokud nepočítáme případ opakovaných lexému v jedné větě). Tudíž na rozdíl od jednotek na nižších jazykových úrovních sémantické konstrukty nemusí být disjunktními množinami svých konstituentů, tj. vět.

Protože je sémantický konstrukt novou jednotkou a navíc jednotkou, která se svou podstatou liší od jednotek na jiných jazykových úrovních, považujeme za nutné, aby byl krátce nastíněn postup kvantifikace textu pro získání dat týkajících se sémantických konstruktů, viz kapitola 4.2. Sémantický konstrukt se vyskytuje prozatím pouze jako konstrukt v nejvyšším binarismu: sémantický konstrukt (v počtu svých vět/klauzí) – věty/klauze (měřené v průměrném počtu svých slov). Není tedy viděn ze dvou úhlů pohledu (jednou jako konstrukt nadřazený konstituentům na nižší úrovni a podruhé jako konstituent podřazený konstruktu na úrovni vyšší) jako většina ostatních jednotek. Dle Hřebíčkovy definice tedy suma všech vět, které obsahují určitou lexikální jednotku, tvoří jeden sémantický konstrukt příslušný dané lexikální jednotce.[18]

4.1.3 Krok 3 – test reprezentativnosti výběrového souboru

Test reprezentativnosti akcentuje tu okolnost, že když vzorek zvětšíme, nic se pro základní soubor signifikantně nezmění, proto je důležité stanovit alespoň rámcově velikost výběrového souboru tak, aby při zvolené směrodatné odchylce odhadů vykazoval reprezentativnost. Při tomto testu dle Kubáčka, viz (Kubáček, 1994), vycházíme z konečného inventáře entit, který má *k* elementů (fonémů, slabik, morfémů, slov, délek vět apod.), *k* je konečné. Z daného materiálu připraveného k analýze vytvoříme předběžný výběr jednotek tak, že pravděpodobnost výskytu každé entity je větší než nula. Máme tedy *k* entit, ke každé přiřazenu příslušnou pravděpodobnost $p_i, p_i > 0$, pro $i = 1, \dots, k$. Na základě získaných dat a formule

$$N = \frac{1}{r^2} \prod_{i=1}^{k} p_i^{\frac{1}{k-1}}, \tag{29}$$

kde p_i je relativní frekvence výskytu jednotlivých entit, *r* je průměrná směrodatná odchylka odhadů, která je dopředu určená a *k* je počet entit v inventáři, obdržíme velikost reprezentativního výběru.

V (Kubáček, 1994) je navrženo, aby vzorec (29) byl upraven logaritmizací a dalšími úpravami.

$$\ln N = \ln \frac{1}{r^2} \cdot p_1^{\frac{1}{k-1}} \cdot \dots \cdot p_k^{\frac{1}{k-1}} = \ln \left[\frac{1}{r^2} \cdot (p_1 \cdot \dots \cdot p_k)^{\frac{1}{k-1}} \right]$$

[18] Pokud se jedna lexikální jednotka opakuje několikrát v jedné větě/klauzi, pak je doporučeno, aby byla počítána pouze jednou, viz též (Hřebíček, 1997), (Hřebíček, 2002) a (Wimmer et al., 2003).

$$\ln N = \frac{1}{k-1} \cdot \sum_{i=1}^{k} \ln p_i - 2 \cdot \ln r \qquad (30)$$

4.1.4 Krok 4 – kvantifikace textů

Na základě stanovení jednotek výše popsanými přístupy jsou kvantifikovány texty a získány tabulky odrážející výsledky na rovinách všech třech zmíněných binarismů obsahující konstrukty s délkami x_i, jejich frekvence z_i a konstituenty s délkami y_i, pro každý binarismus označený $i = 1,2,3$.

4.1.5 Krok 5 – výpočet parametrů $A_i, b_i, c_i, i = 1, 2, 3$

V této kapitole bude především nastíněn způsob, který umožňuje odhadnout parametry A_i, b_i, c_i a následně vypočítat převrácené hodnoty jejich aritmetického průměru *D* nejprve pomocí statistických metod, konkrétně technikou lineární regrese, viz např. (Kubáček & Kubáčková, 2000). Úkolem je najít regresní křivku, v našem případě regresní přímku, která co nejlepším způsobem aproximuje logaritmicky transformovaný lineární model. Následně je model testován na svou spolehlivost taktéž metodami statistickými. V další sekci této kapitoly je popsána alternativní metoda výpočtu parametrů pomocí numerické analýzy, konkrétně Gauss-Newtonovým algoritmem. Nevýhodou tohoto způsobu výpočtu, který je přesnější než metoda regresní, je, že neumožňuje testovat model na spolehlivost.

4.1.5.1 Výpočet pomocí statistických metod

V této kapitole je nutné jako další krok popsat způsob, jakým je možné vypočítat koeficienty A_i a b_i, c_i, $i = 1,2,3$. Tyto koeficienty jsou provázány prostřednictvím formule Menzerath-Altmannova zákona, ať ve své zkrácené, nebo v úplné verzi. Proměnné x_i (nezávislá proměnná) a y_i (závislá proměnná) jsou obě numerické (kvantitativní) proměnné.

Jednou z technik, kterou je možné použít pro získání všech koeficientů a která v tomto experimentu byla použita, je metoda lineární regrese. Metoda byla aplikována na zkrácenou (20) i úplnou formuli (21) Menzerath-Altmannova zákona. Podrobný postup si ukážeme na metodě lineární regrese, která je svou povahou jednodušší než nelineární regrese. Účelem lineární regrese je nastínit vztah mezi oběma numerickými proměnnými tak, že daný vztah lze vyjádřit matematickou rovnicí. Charakter tohoto vztahu je dán právě parametry A_i, b_i a c_i[19]. V našem experimentu je situace modifikovaná tím, že vztah (MAL) je předem znám a lineární regresi používáme výhradně k tomu, abychom našli koeficienty A_i a b_i, $i = 1,2,3$ a posléze zjistili, jak těsný je vztah mezi nimi (viz následující kapitola). Existují dvě formy vztahu mezi statistickými proměnnými. První z nich je vztah funkční a druhý statistický. První znamená, že jedna hodnota první proměnné koresponduje s jednou hodnotou proměnné druhé a naopak. Statistický vztah vyjadřuje, že existuje více hodnot druhé proměnné korespondujících s jednou hodnotou proměnné první, tj. kvůli změnám hodnot jedné proměnné se také mění pravděpodobnostní rozdělení změn proměnné druhé, viz např. (Svatošová & Kába, 2009). Je nutné poznamenat, že čím jednodušší je tento vztah (tudíž i matematická formule), tím lépe. Tudíž pokud je to proveditelné a vhodné, volíme regresi lineární. Ale abychom tak mohli učinit,

[19] Za běžných okolností se regrese používá právě pro nalezení vztahu mezi hodnotami empirických pozorování.

je třeba transformovat zkrácený vztah MAL (20) nebo jeho úplnou variantu (21), aby vyjadřovaly lineární vztah. Logaritmujeme tedy celou rovnici (je libovolné, jaký základ logaritmu si zvolíme) a dostáváme pro $i = 1, 2, 3$

$$\ln y_i = \ln A_i - b_i \cdot \ln x_i \tag{31}$$

$$\ln y_i = \ln A_i - b_i \cdot \ln x_i - c_i \cdot x_i. \tag{32}$$

Každá z tabulek obsahujících výstupy kvantifikace z předchozího kroku formuje sekvenci n_i datových bodů, které jak předpokládáme, vyhovují transformovaným výše zmíněným rovnicím, ke kterým přičítáme normálně rozdělené chyby ε_i^j, kde $i = 1, 2, 3, j = 1, 2, \dots, n_i$, tedy např. ($Y_i^j$ označuje náhodnou proměnnou)

$$\ln Y_i^j = \ln A_i \ - b_i \cdot \ln x_i^j + \varepsilon_i^j, \ i = 1{,}2{,}3, j = 1{,}2, \dots, n_i, \tag{33}$$

$$\ln Y_i^j = \ln A_i \ - b_i \cdot \ln x_i^j + c_i x_i^j + \varepsilon_i^j, \ i = 1{,}2{,}3, j = 1{,}2, \dots, n_i. \tag{34}$$

Obecně pro $i = 1, 2, 3$ mluvíme o lineárním modelu (modelu jednoduché regrese)

$$\boldsymbol{Y} \sim N_{n_i}(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}), \tag{35}$$

kde

$$\boldsymbol{Y} = \begin{pmatrix} \ln Y_i^1 \\ \vdots \\ \ln Y_i^{n_i} \end{pmatrix}$$

a

$$\boldsymbol{X} = \begin{pmatrix} 1 & \ln x_i^1 \\ \vdots & \vdots \\ 1 & \ln x_i^{n_i} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \ln A_i \\ -b_i \end{pmatrix}$$

(model zkráceného tvaru formule MAL odpovídajícímu rovnici (20)) nebo

$$\boldsymbol{X} = \begin{pmatrix} 1 & \ln x_i^1 & x_i^1 \\ \vdots & \vdots & \vdots \\ 1 & \ln x_i^{n_i} & x_i^{n_i} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \ln A_i \\ -b_i \\ c_i \end{pmatrix}$$

(model úplného tvaru formule MAL odpovídající rovnici (21)).

Aby byla skutečně na první pohled patrná proklamovaná linearita, provedeme u rovnice (31) zkráceného tvaru formule MAL substituci $y' = \ln y, x' = -\ln x$ a $a = \ln A$. Takto konečně dostáváme z (31)

$$y' = a + b \cdot x', \tag{36}$$

což je zjevně rovnice, jejíž grafickou reprezentací je přímka. Absolutním členem rovnice této přímky je koeficient a a směrnicí je koeficient b.

Nejdříve tedy potřebujeme transformovat původní proměnné x_{ij} a y_{ij}, kde $i = 1,2,3$ jsou prověřované lingvistické úrovně a $j = 1, \cdots, n, n \in N$ jsou všechna empirická pozorování na každé z příslušných jazykových úrovní.

Lineární regrese ukazuje vztah mezi oběma numerickými proměnnými x_{ij} a y_{ij}, kde $i = 1,2,3$, tak, že determinuje přímku, která aproximuje co nejlépe body, které jsou grafickou reprezentací jednotlivých pozorování v bodovém grafu, viz (Petrie & Watson, 2006). V našem případě můžeme použít *jednoduchou (jednorozměrnou) lineární regresi*, protože máme v našem experimentu pouze jednu nezávislou proměnnou. Pokud je vztah skutečně lineární, jak bylo požadováno výše, pak může být následně graficky zaznamenán jako přímka aproximující vztah mezi oběma proměnnými. Samozřejmě může být přímka načrtnuta od oka v bodovém grafu, ale nepřesnost, která by takto mohla vzniknout, není žádoucí. Abychom zaručili maximální přesnost, vybudujeme matematický model reprezentující reálnou situaci nebo proces, který se objevuje v základním souboru, viz (Petrie & Watson, 2006).

Naším úkolem je získat odhady a'_i, b'_i parametrů $a_i, b_i, i = 1,2,3$, v odpovídajícím pořadí z našeho náhodného vzorku *n* párů $\{(x'_1, y'_1), (x'_2, y'_2), \cdots, (x'_n, y'_n)\}$ na každé z našich třech zkoumaných jednotlivých jazykových úrovní. Koeficienty a'_i, b'_i se nazývají *regresní koeficienty*. Abychom poté získali zpět původní koeficient *A*, je nutné odlogaritmovat původní substituci následujícím způsobem

$$a = \ln A \Leftrightarrow A = e^a. \tag{37}$$

Mezi dalšími požadavky týkajícími se zmíněné regresní přímky je, že by měla být umístěna co nejblíže, jak je možné, k bodům v bodovém grafu, tzn., odchylka přímky od bodů by měla být co nejmenší. Takovéto odchylky jsou vertikálními vzdálenostmi bodů od regresní přímky a nazývají se *rezidua*. Pokud jsou body umístěny nad přímkou, pak hodnota odpovídajících reziduí nabývá kladných hodnot, pokud jsou pod přímkou, pak jsou rezidua záporná. Abychom tedy předešly možným problémům se znaménky reziduí, budeme minimalizovat sumu kvadrátů reziduí, abychom získali koeficienty a'_i, b'_i. Tento způsob se nazývá *metoda nejmenších čtverců*. Pomocí ní můžeme odhadnout parametry $\boldsymbol{\beta}$, viz (Ralston, 1965) a (Stoer & Bulirsch, 2002). Takto mohou být proměnné a koeficienty regresních rovnic pro každé $i = 1,2,3$

$$y'_i = a'_i + b'_i \cdot x'_i \tag{38}$$

Interpretovány následujícím způsobem: y'_i pro každé $i = 1,2,3$ jsou logaritmy průměrných hodnot $y_i, j = 1,2, \cdots, n, a'_i$ jsou odhady absolutních členů odpovídajících přímek.

Úkolem lineární regrese (a potažmo nyní i úkolem naším) je najít a zanést do grafu přímku, která nejlépe aproximuje body, které korespondují s logaritmy našich pozorovaných hodnot tak, že rezidua jsou minimální. Tudíž je pravda, že na všech zkoumaných lingvistických úrovních, pro všechna $j = 1,2, \cdots, n$ platí, že

$$\sum_{j=1}^{n} \left(y'_j - a - b \cdot x'_j\right)^2 \rightarrow \min \ , \tag{39}$$

tedy že suma všech kvadrátů příslušných reziduí je minimální. Odsud po úpravách dostáváme vzorce pro výpočet koeficientů a', b' (odhady koeficientů a, b)

$$b = -\frac{n\cdot\sum_{j=1}^{n} x'_j y'_j - \sum_{j=1}^{n} x'_j \sum_{j=1}^{n} y'_j}{n\cdot\sum_{j=1}^{n} x'^2_j - \left(\sum_{j=1}^{n} x'_j\right)^2} \tag{40}$$

$$a = \frac{\sum_{j=1}^{n} x'^2_j \sum_{j=1}^{n} y'_j - \sum_{j=1}^{n} x'_j y'_j \sum_{j=1}^{n} x'_j}{n\cdot\sum_{j=1}^{n} x'^2_j - \left(\sum_{j=1}^{n} x'_j\right)^2}, \text{ viz (Wimmer et al., 2003).} \tag{41}$$

Tudíž regresní rovnice je

$$y' = a + b \cdot x'. \tag{42}$$

Přímka, která je reprezentovaná regresní funkcí (42), je nejlepším odhadem teoretické regresní přímky, která má rovnici

$$Y'_{pop} = \alpha' + \beta' x', \tag{43}$$

kde α'je logaritmus absolutního členu rovnice teoretické regresní přímky a β'je logaritmus směrnice teoretické regresní přímky; a', b' jsou nestranné, konsistentní a dostatečné odhady parametrů α'a β'.

Předpokládejme, že $y'_i, i = 1,2, \dots n$ jsou logaritmy empirických nebo pozorovaných hodnot proměnné *Y*, a y''_i, $i = 1,2, \dots n$ jsou logaritmy teoretických hodnot získaných z regresní rovnice (42), pak

$$d'_i = y'_i - y''_i \tag{44}$$

se nazývají rezidua, jak již bylo definováno výše. Následně,

$$S_r = \sum_{i=1}^{n} d'^2_i \tag{45}$$

je reziduální součet čtverců, a

$$s_r^2 = \frac{S_r}{n-2} = \frac{\sum_{i=1}^{n}(y'_i - a - bx'_i)^2}{n-2} = \frac{\sum_{i=1}^{n} y'^2_i - a\sum_{i=1}^{n} y'_i - b\sum_{i=1}^{n} x'_i y'_i}{n-2} \tag{46}$$

je *reziduální rozptyl*, který vyjadřuje, jak mnoho jsou hodnoty proměnné *Y'* rozptýleny kolem regresní funkce. Reziduální odchylka je začleněna do formule použité pro testování hypotézy a parametrech přímky a pro výpočet konfidencích intervalů. Tento proces bude demonstrován v další kapitole, viz také (Petrie & Watson, 2006).

Na závěr považuji za nutné zmínit se o důvodech, proč se zabývat lineární regresí. Za prvé, chceme odhalit, zda existuje kauzální vztah mezi studovanými proměnnými. Za druhé, regresní funkce nám umožňuje předpověď dalšího vývoje založenou na regresních odhadech, což znamená, že dokážeme předpovědět hodnoty závislé proměnné ze známých nebo předpokládaných hodnot proměnné nezávislé. A konečně speciálně pro tento experiment používáme regresní analýzu proto, abychom nalezli koeficienty *A* a *b* a abychom testovali spolehlivost výsledků tohoto experimentu.

4.1.5.2 *Výpočet numerickými metodami*

Pokud se týká spolehlivosti experimentu, logaritmická transformace a lineární regrese nám neposkytují naprosto spolehlivá data týkající se požadovaných parametrů b_i, i = 1, 2, 3, protože konfidenční intervaly jsou příliš široké a některé obsahují též hodnotu nula, detaily viz v praktické části. Ale naštěstí existuje ještě jeden způsob, jak najít parametr u rovnic (15) a (16). Je možné použít *metody numerické*.

V případě numerických metod je ještě více než dříve doporučeno využít služeb statistického softwaru, detaily viz kapitola 4.2.5.2.

4.1.6 Krok 6 – statistická analýza

Testování hypotézy o parametrech lineární regrese

Před tím, než budeme testovat hypotézy o parametrech lineární regrese, musíme být velice opatrní a prověřit určité, přesně dané předpoklady, které tvoří podklad pro lineární regresi. Jedná se o následující předpoklady:

- Vztah mezi proměnnými *x'* a *y'* je lineární.
- Proměnná *x'* je měřena bez chyby.
- Pro každou hodnotu proměnné *x'* mají hodnoty závislé proměnné *y'*, ze které vybíráme náš výběrový soubor, normální rozdělení.
- Pro každou hodnotu proměnné *x'* leží průměrná hodnota rozdělení hodnot základního souboru proměnné *y'* na přímce reprezentované rovnicí (15), (16).
- Rozptyl rozdělení hodnot základního souboru hodnot proměnné *y'* je konstantní pro každou hodnotu proměnné *x'.*
- Pozorování jsou nezávislá. Pro další detaily viz (Petrie & Watson, 2006).

Pro takováto testování můžeme použít výše zavedených reziduí například způsobem, který v krátkosti naznačíme v následujících bodech:

- Sestrojíme si graf závislosti reziduí (44) na hodnotách proměnné *x'*. Pokud je vztah mezi proměnnými *x'* a *y'* lineární, jsou rezidua rozptýlena kolem nuly. Není patrný žádný rostoucí ani klesající trend.
- Abychom ověřili normální rozdělení reziduí, je vhodné si sestrojit například histogram reziduí.
- Abychom prověřili rozptyl, sestrojíme si graf závislosti reziduí na odpovídajících (předpokládaných) hodnotách. V případě, že jsou rezidua náhodně rozptýlena, je předpoklad, že je rozptyl reziduí konstantní, splněn. Pokud je z grafu patrná jakákoli tendence hodnot, například konická nebo parabolická, rozptylu reziduí klesat nebo stoupat, pak není předpoklad splněn. Pro další detaily viz (Petrie & Watson, 2006).

Následně, po prověření předpokladů, můžeme shrnout regresní diagnostiku do následujících kroků.

1. Nejprve musíme obecně specifikovat *nulovou hypotézu* $H_0: \beta = 0$ (To znamená, že teoretický regresní koeficient je nula, tedy že neexistuje lineární závislost mezi oběma proměnnými.) a *alternativní hypotézu* $H_1: \beta \neq 0$.

2. Jelikož jsme získali nejlepší možnou regresní přímku (viz výše) reprezentovanou rovnicí (15) či (16), můžeme prověřit předpoklady pro splnění lineární regrese pomocí zkoumání vlastností reziduí, jak již bylo popsáno v předcházejícím odstavci.
3. V následujícím kroku spočítáme odpovídající *testové kritérium*. Je možné tak učinit s pomocí vhodného softwaru, nebo mechanicky pomocí následujících formulí

$$t = \frac{b}{s_b} = \frac{b}{s_r}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{47}$$

$$\text{pro } s_b = \sqrt{\frac{s_r^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \tag{48}$$

   kde s_r je směrodatná odchylka reziduí, $\bar{x}$ je průměrná hodnota výběru. Kritérium splňuje *t*-rozdělení a má *n*-2 stupně volnosti.
4. Kroky 3. a 4. mohou být a obvykle jsou zpracovány pomocí vhodného softwaru, například SAS, Statistica a R. Typický příklad výstupu zpracování pomocí takového softwaru bude uveden v praktické části této kapitoly a jeho statistická část bude dále komentována.
5. Nyní musíme rozhodnout, zda zamítnout nulovou hypotézu, nebo ne. Obvykle zamítáme nulovou hypotézu v případě, že $P < 0{,}05$. Pokud zamítneme hypotézu, že $H_0: \beta = 0$, pak můžeme říct, že regresní koeficient *b* je *statisticky významný*. V takovém případě jsme oprávněni použít rovnici regresní přímky (15) pro zpracování regresních odhadů.
6. V souvislosti s obecným algoritmem odhadu intervalu můžeme zkonstruovat *interval spolehlivosti*[20] pro teoretický koeficient *β*. Můžeme jej vypočítat opět mechanicky, pomocí formule

$$\left(b - t_{\alpha(n-2)}.s_b, b + t_{\alpha(n-2)}.s_b\right), \tag{49}$$

   kde $t_{\alpha(n-2)}$, pro $\alpha = 0{,}05$ je kritická hodnota získaný z tabulky *t*-rozdělení s *n*-2 stupni volnosti. Pro další detaily viz např. (Petrie & Watson, 2006) a (Svatošová & Kába, 2009). I k výpočtu intervalů spolehlivosti je efektivnější použít statistický software, výstupy využitého R softwaru pro náš experiment budou uvedeny dále.

Lineární korelace

V předcházejících krocích naší analýzy bylo naším cílem zjistit formu závislosti a vyjádřit ji matematicky takzvanou regresní funkcí (viz výše). V následujícím kroku našeho algoritmu budeme zkoumat stupeň intenzity, se kterou se daná závislost objevuje mezi ostatními

[20] Velice častým úkolem statistiky je na základě daných dat odhadnout příslušný parametr. Odhadujeme-li tento parametr jedním číslem, pak je bodovým odhadem, tzn., je zatížen chybou (protože je založen na náhodně posbíraných datech, která mohou být vychýlená). Bodový odhad se tedy týká výběrového souboru. Chceme-li ale odhad rozšířit na celou populaci, je lépe používat odhad intervalový. Hodnota spolehlivosti $(1 - \alpha)$ udává pravděpodobnost, s níž je skutečná hodnota parametru nalezeným intervalem pokryta.

rušivými faktory. Takovéto zkoumání je úkolem pro lineární korelaci. Jinými slovy, lineární korelace měří, jak dobře popisuje přímka (křivka lineární regrese) lineární vztah mezi dvěma proměnnými.

Výchozí bod pro měření síly závislosti je daný regresní model. Základní důvody pro měření síly závislosti jsou následující:

- Čím silnější je vztah mezi dvěma proměnnými, tím víc můžeme očekávat, že změny jedné proměnné způsobí změny proměnné druhé.
- V naší analýze je ještě mnohem zásadnější zjistit vysvětlující sílu použitého regresního modelu nebo dalších předcházejících kroků v našem algoritmu. Čím menší je rozptýlení empirických hodnot závislé proměnné kolem odpovídající regresní křivky (to znamená, čím je závislost silnější), tím přesnější budou regresní odhady založené na dané regresní funkci, viz (Svatošová & Kába, 2009). Jinými slovy tím přesněji v našem experimentu určíme parametry A_i, b_i.

Existují jisté způsoby, jak změřit sílu závislosti. Jeden z nejcharakterističtějších, který si velice stručně popíšeme, je použití *korelačního koeficientu* (*Pearsonova korelačního koeficientu*), který vyjadřuje rozpětí, ve kterém jsou body rozptýleny kolem přímky. Nabývá hodnot intervalu $\langle -1,1 \rangle$. Pokud existuje lineární korelace mezi dvěma proměnnými (body $[x'_1, y'_1], [x'_2, y'_2], \dots, [x'_n, y'_n]$ se nacházejí na přímce), pak $|r| = 1$. Na druhé straně, pokud proměnné závislé vůbec nejsou (jsou nekorelované), pak $r = 0$. Pro další detaily viz např. (Petrie & Watson, 2006) a (Svatošová & Kába, 2009).

Jestliže zkoumáme náhodný vzorek *n* pozorování $\{(x'_1, y'_1), (x'_2, y'_2), \dots, (x'_n, y'_n)\}$ dvou numerických proměnných, pak můžeme odhadnout korelační koeficient ρ v populaci pomocí korelačního koeficientu výběrového souboru

$$r = \frac{\sum_{i=1}^{n}(x'_i - \bar{x})(y'_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x'_i - \bar{x})^2 \sum_{i=1}^{n}(y'_i - \bar{y})^2}}, \tag{50}$$

který je platný pouze v mezích dat ve výběrovém souboru, viz (Petrie & Watson, 2006). S ohledem na tyto vlastnosti, uvedu následující pomůcku, která je ale pouze konvenční a nikoliv zavazující a která slouží k odhadu síly lineární korelace mezi dvěma proměnnými:

$0 < \|r\| \leq 0{,}3$	mírná závislost
$0{,}3 < \|r\| \leq 0{,}8$	středně silná závislost
$0{,}8 < \|r\| \leq 1$	silná závislost. [Sv]

Shrnutí

Shrňme si tedy v několika jednoduše formulovaných bodech postup statistické analýzy:

1. Nejprve na základě hodnot obou sad proměnných získaných kvantifikací textu určíme parametry $A_i, b_i, i = 1,2,3$ a získáme regresní analýzou rovnici regresní přímky $y' = a + b \cdot x'$.

2. Stanovíme konfidenční interval pro regresní koeficient β, nejčastěji 95% konfidenční interval.
3. Zjistíme sílu závislosti mezi proměnnými.
4. Na hladině významnosti $\alpha = 0{,}05$ provedeme test významnosti korelačního koeficientu *r*. Pro další detaily viz (Svatošová & Kába, 2009).

Koeficient determinace

V následující části se pokusím nastínit další možný způsob verifikace spolehlivosti zvoleného modelu, tj. způsob, jakým si můžeme spočítat, jak těsně přiléhá regresní přímka k izolovaným bodům znázorňujícím naše pozorování. Jako měřítko této těsnosti je používán *koeficient determinace* R^2. Koeficient determinace tedy zjišťuje adekvátnost přiřazení funkce k empirickým datům.

Koeficient determinace může být vypočten mechanicky dle vzorce, viz např. (Wimmer et al., 2003),

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_j - y_{j\,\mathrm{est}}\right)^2}{\sum_{i=1}^{n}\left(y_j - \bar{y}\right)^2}, \tag{52}$$

kde y_j jsou empirická data, $y_{j\,\mathrm{est}}$ jsou data získaná výpočtem z MAL při použití získaných parametrů A_i, b_i, popř. c_i, $\bar{y}$ je aritmetický průměr y_j.

4.1.7 Krok 7 – fraktální analýza

Snadno můžeme vyjádřit výše několikrát zmíněnou úplnou indexovanou verzi formule MAL na *n* = 3 lingvistických úrovních (21), tj.

$$y_i = A_i . x_i^{-b_i} . e^{c_i x_i}, i = 1{,}2{,}3,$$

Může být ekvivalentně vyjádřena jako

$$\frac{1}{b_i} = \frac{\log x_i}{\log\left(\frac{A_i}{y_i} . e^{c_i x_i}\right)} = \frac{\ln x_i}{\ln\left(\frac{A_i}{y_i} . e^{c_i x_i}\right)}, i = 1{,}2{,}3.$$

Její jednoduchá varianta (20) pro $c_i = 0$, tj.

$$y_i = A_i . x_i^{-b_i}, i = 1{,}2{,}3,$$

dostává ekvivalentní formu

$$\frac{1}{b_i} = \frac{\log x_i}{\log \frac{A_i}{y_i}}, i = 1{,}2{,}3.$$

Pro další detaily tohoto jednoduchého, ale velice důležitého pozorování viz (Hřebíček, 2000) a (Hřebíček, 2007).

Toto nám v pohledu známé *Moran-Hutchinsonovy formule* pro výpočet fraktální dimenze *D*, umožňuje *interpretovat převrácenou hodnotu aritmetického průměru* $\frac{3}{b_1+b_2+b_3}$ *koeficientů* b_1, b_2, b_3 *jako dimenzi* $D = \dim(\mathbf{A})$ *vhodného cyklicky soběpodobného fraktálu* **A**, pro další detaily viz (Andres, 2009) a (Andres & Rypka, 2011), tj.

$$D := \frac{3}{b_1 + b_2 + b_3}.$$

$$\text{Pro } x := x_1 = x_2 = x_3 \text{ a } r_i := \left(\frac{y_i}{A_i.e^{c_i x_i}}\right)^k = \frac{1}{x^{kb_i}}, i = 1,2,3,$$

$$\left(\Rightarrow r_1 r_2 r_3 := \left(\frac{y_1 y_2 y_3}{A_1 A_2 A_3.\, e^{(c_1+c_2+c_3)x}}\right)^k = \frac{1}{x^{k(b_1+b_2+b_3)}}\right),$$

kde nutně pro každé $i = 1,2,3 \max\frac{1}{b_i} \leq k \in N$, může být fraktál **A** považován za jedinečnou uzavřenou pozitivně invariantní množinu $\mathbf{A} = F(\mathbf{A})$ složeného zobrazení $F = F_3 \circ F_2 \circ F_1$ *Hutchinson-Barnsleyho zobrazení* F_i, kde

$$F_i(\boldsymbol{x}) := \bigcup_j {}_i f_j(x), \; {}_i f_j : [0,1]^k \to [0,1]^k, \tag{53}$$

$${}_i f_{\boldsymbol{j}}(\boldsymbol{x}) := r_i \boldsymbol{x} + \frac{1}{x}\boldsymbol{j}, \boldsymbol{j} = (j_1, \dots, j_k), j_l \in \{0,1, \dots, x-1\}, i = 1,2,3.$$

Dále může být získán jako limitní množina (s odkazem na Hausdorffovu metriku d_H) postupných aproximací $F^0([0,1]) := [0,1], F^s([0,1]), s = 1,2, \dots,$ **A**, tj. $\lim_{s\to\infty} d_H(F^s([0,1]), \mathbf{A}) = 0$, kde Hausdorffova vzdálenost $d_H\left(F^S([0,1], \mathbf{A})\right)$ mezi aproximacemi a **A** může být odhadnuta následujícím způsobem:

$$d_H = \left(F^s([0.1], \mathbf{A})\right) \leq \frac{(r_1 r_2 r_3)^s}{1 - r_1 r_2 r_3} d_H\left([0,1], F([0,1])\right) =$$

$$= \left(\left(1 - \frac{1}{x}\right) + \left(1 - \frac{1}{x}\right) r_2 r_3\right)\sqrt{k-1} / \left(x^{sk(b_1+b_2+b_3)}\left(1 - x^{-k(b_1+b_2+b_3)}\right)\right) \leq \frac{(r_1 r_2 r_3)^s}{1 - r_1 r_2 r_3}\sqrt{k}. \tag{54}$$

Povšimněme si, že pro $A := A_1 = A_2 = A_3, b := b_1 = b_2 = b_3$ a $c := c_1 = c_2 = c_3$ hodnota $\frac{1}{b}$ *může být jednoduše interpretována jako fraktální dimenze fraktálu* $\mathbf{A} = F_1(\mathbf{A}) = F_2(\mathbf{A}) = F_3(\mathbf{A})$, protože, z pohledu výše zmíněné shody, máme $r := r_1 = r_2 = r_3 = \left(\frac{y}{A.e^{cx}}\right)^k =$

$\frac{1}{x^{kb}} \leq \frac{1}{x}$. Redukovaná formule $(c = 0)$ pak vyžaduje pouze, aby bylo položeno $r := \left(\frac{y}{A}\right)^k = \frac{1}{x^{kb}}$.

Fraktální dimenze $D^{(p)}$ *p*-dimenzionální projekce **A** může být spočtena jako $D^{(p)} = \frac{p}{k} D$.

Pro další detaily týkající se teoretických aspektů fraktální analýzy viz (Andres, 2009), (Andres & Rypka, 2011) a (Barnsley, 1988).

4.1.8 Krok 8 – vizualizace

4.1.8.1 Vizualizace fraktálem

Z pohledu výše zmíněné fraktální analýzy může být složené zobrazení

$$A_1 := F_1([0,1]), A_2 := F_2 \circ F_1([0,1]), A_3 := F_3 \circ F_2 \circ F_1([0,1]) = F([0,1])$$

považováno, dle výše zmíněné shody, za vizualizovanou strukturu lingvistických objektů na *n = 3* lingvistických úrovních charakterizovaných koeficienty A_i, b_i, c_i (*i* = 1,2,3) na MAL.

Všimněme si, že pro $A_{s3} := F^s([0,1])$ máme podle výše zmíněných argumetů $\lim_{s\to\infty} d_H(A_{s3}, \mathbf{A})$ a výše zmíněný odhad pro Hausdorffovu vzdálenost $d_H(A_{s3}, \mathbf{A})$ mezi A_{s3} a **A** vyhovuje.

Navíce se F^s skládá z x^{3ks} kontrakcí se stejným faktorem $r := x^{-k(b_1+b_2+b_3)}$.

Pro vizualizace výše zmíněného složeného zobrazení A_1, A_2, A_3 a množin $A_{s3}, s = 1,2, \ldots$, pro danou úvodní množinu [0,1], využijeme tu nejposlednější iteraci. Úvodní množina nijak neovlivní výstupní atraktor, ale může být důležitá pro grafické znázornění iterací. Pro simplifikaci je výhodné determinovat jednoduché množiny několika body. V našem případě byly použity úsečky, které jsou definované dvěma body. Dosazením do vzorců můžeme spočítat souřadnice bodů (obrazů), jejichž počet je x^k-krát násobný. V *s*-tém kroku dostaneme $2x^{3ks}$ bodů. Tímto způsobem jsme schopni spočítat pouze několik iterací, ale obvykle jsou po několika krocích následující iterace nerozlišitelné. Délka úseček v *s*-tém kroku je

$$\prod_{k=1}^{3s} r_{i_k}, i_k \in \{1,2,3\}. \quad (55)$$

Když takto obdržíme dvojice jednotlivých bodů, jsme díky nim snadno schopni graficky znázornit úsečky, které jsou posledními iteracemi. Kvůli rozlišení monitorů a oka nemá smysl provádět kontrakce úseček kratších než tisíciny znázorněné délky intervalu.

V našem případě uvažujeme složené zobrazení $F = F_3 \circ F_2 \circ F_1$ tří Hutchinson-Barnsleyho zobrazení ve vzorci (53) a jeho projekce do dvoudimenzionálního prostoru, tj. bereme x^2 podobností. Je třeba poznamentat, že vytvoření jednoho systému složením *n* = 3 zobrazení F_1, F_2, F_3 by bylo proveditelné a obsahovalo by x^6 zobrazení. Nicméně možnost modelovat segmentaci jazykových struktur by byla ztracena. Jakékoli složení kontrakcí

(podobností) je opět kontrakce (podobnost), tj. existuje atraktor složeného zobrazení *F* a iterace původní množiny úseček se bude opět skládat z úseček. Tudíž vytvoříme posloupnost

$$[0,1], F_1([0,1]), F_2\big(F_1([0,1])\big), F_3\left(F_2\big(F_1([0,1])\big)\right) = F([0,1]), F_1\big(F([0,1])\big), F_2\left(F_1\big(F([0,1])\big)\right),$$

$$F^2([0,1]) = F\big(F([0,1])\big), F_1(F^2[0,1]), \ldots$$

Ale graficky znázorníme výhradně iterace složeného zobrazení $F^s([0,1])$.

Iterace úseček je velice snadno možné vykreslit v MATLABu. Jak již bylo poukázáno, stačí znát pouze koncové body úseček vzniklých zobrazením v (53), protože příkaz `line` v MATLAB spojuje oba koncové body úseček.

4.1.8.2 Shluková analýza

Termín *shluková analýza* se používá pro označení široké škály logických výpočetních postupů, kterými můžeme objektivně shlukovat jedince do relativně homogenních podmnožin – *shluků* – podle jejich podobností nebo naopak dle rozdílů mezi nimi. Rozklad by měl být prováděn tak, že objekty, které se nachází uvnitř jednotlivých shluků, jsou si, jak jen je to možné, podobné. Na druhou stranu objekty, které náleží do shluků různých, jsou si podobné co možná nejméně. Pro naši analýzu použijeme aglomerativní přístup, což je jedna z hierarchických metod shlukové analýzy.

Funkce shlukové analýzy jsou následující. Shluková analýza umožňuje analyzovat, zda se množina objektů přirozeně rozpadá na jednotlivé podmnožiny (shluky) objektů podobných jeden druhému uvnitř shluku a zároveň odlišných od objektů náležejících do shluků ostatních. Dále je možné zjistit, zda existuje celá hierarchie takových rozkladů. Pokud dále existují nějaké shluky, pak je možné metodami shlukové analýzy odhalit jejich vlastnosti. Shluková analýza též umožňuje zjistit způsob, jakým se další potenciální objekty integrují do již existujících shluků.

Jednotlivé kroky algoritmu shlukové analýzy jsou následující:

1. Výpočet matice podobností objektů. Úvodní rozklad je tvořen shluky, které jsou tvořeny jedním objektem.
2. Nalezení nejmenší vzdálenosti mezi jednotlivými shluky na dané úrovni hierarchie.
3. Sdružení nejbližších shluků do jednoho shluku společného na nejbližší vyšší úrovni hierarchie. Ostatní shluky zůstanou nezměněny.
4. Výpočet charakteristik shluků na dané úrovni hierarchie.
5. Pokud stále zůstává více než jeden shluk, celý algoritmus je nutné zopakovat.

Pro další podrobnější informace o shlukové analýze viz např. (Jain & Dubes, 1998).

Pro vykreslování shlukovacích tendencí jsou používány *dendrogramy*. Tento specifický typ stromových diagramů dokáže velice efektivně demonstrovat vztahy mezi jednotlivými shluky. Může též znázorňovat vícerozměrné vzdálenosti mezi objekty. Nejbližší shluky nebo objekty jsou spojovány horizontálními čarami.

4.1.9 Krok 9 – interpretace získaných výsledků analýzy

Posledním velice důležitým krokem celého algoritmu je interpretace získaných kvantitativních dat, obrázků a grafů. Konkrétní interpretace v případě výběrových souborů zvolených pro tento experiment je zařazena na závěr následující praktické části.

4.2 Praktická aplikace algoritmu kvantitativní analýzy textu

4.2.1 Krok 1 – volba výběrového souboru

Havran je chemickou sloučeninou poezie a matematiky.

Olla Hanson

V roce 1845 vznikla v New Yorku báseň *The Raven.* Jejím autorem byl Edgar Allan Poe, který v roce 1846 doplnil *Havrana* o *Filozofii básnické skladby (The Philosophy of Composition)*, ve které vysvětluje, jak elegantním způsobem zcela vědomě docílil ponuré atmosféry a tísně, která dopadá na posluchače či čtenáře. Je nepopiratelné, že Poe ukázal velikost svého intelektu nejen v této básni. Ale právě k této básni poskytl „manuál", ve kterém vysvětlil, jak cíleně vznikala. Mimo jiné zmiňuje důvod volby délky, což je čistě kvantitativní kritérium, „budiž rozsah básně v matematickém poměru k její hodnotě – jinými slovy k vzruchu, k povznesení – anebo ještě k stupni pravého básnického účinku, jaký dovede navodit; neboť je jasné, že krátkost musí být v přímém poměru k mohutnosti zamýšleného účinku…" (Poe, 1985, s. 73). Mimo jiné stanoví též „čep, kolem kterého by se mohla celá stavba otáčet" (Poe, 1985, s. 74), což je refrén, který musí splňovat jistá kritéria zvuku i myšlenky a musí navozovat jednotvárnost. Poe volí jedno slovo (opět kvantitativní kritérium) „nevermore", které se bude opakovat na konci drtivé většiny slok pod různými záminkami. Samo slovo „nevermore" vybral Poe pod vlivem zvukového kritéria. Poe též polemizuje o požadavcích uvalených na délku sloky a rýmy[21].

Tento fakt mi vnuknul myšlenku podrobit báseň také kvantitativnímu rozboru. To, že báseň je působivá tak, jak Poe zamýšlel, je nesporné, proto mým záměrem nebylo tento fakt potvrdit, ale spíše vyzkoušet funkčnost a oprávnit další používání výše zmíněných vzorců MAL, definic a postupů.

Text je zvolen jako předmět zkoumání, protože více zachovává pravidla a struktury, dále protože je lépe zachytitelný a uchopitelný než promluva a dá se lépe zpracovat. Je ale na druhé straně tak mnohovrstevnatý a obsahuje tolik elementů, že je poměrně složité vybrat, co a jakým způsobem porovnávat, proto zdůrazňuji, že je tato práce náhledem na několik možných kvantitativních experimentů. Ke zkoumání jsem zvolila Poeův originální text v anglickém jazyce, jeden překlad do německého jazyka a šestnáct českých překladů[22,23] (Poe,

[21] V analýze připojené dále se nepracuje ani s délkou sloky ani s délkou rýmu. Na závěr této práce v diskusi ale připojuji možnosti dalšího zkoumání, kde se ukazuje jako jedna potenciální cesta pro zkoumání akceptace formálních kritérií, tedy i například délky sloky či rýmu.

[22] Poeův *Raven* výrazně zasáhl do české překladatelské tradice. Přibližně za sto let existence básně vzniklo v Čechách kolem dvaceti překladů této básně, z nichž některé podrobím výše zmíněné analýze, viz (Poe, 1985), (Poe, 2008a), (Poe, 2008b). Mnoho dalších překladů vzniklo i poté. Báseň je však obecně

1985). Dostupnost originálního textu a velkého množství překladů, které jsou nuceny více či méně se řídit závaznými pravidly pro formu i obsah a navíc, které je možno najít ve velkém množství jazyků odlišných svou strukturou, je velice silnou motivací pro kvantitativní analýzu.

Tato zdánlivě jednoznačná volba Poeovy básně díky své „matematičnosti" s sebou ale přináší značná úskalí. Poeův *Havran* je textem poetickým, takže se jeho stavba vždy neřídí striktně jazykovými pravidly, ale naopak je velice často záměrně porušuje. O úskalích této volby blíže viz kapitola pojednávající o volbě jednotek. Frekvence jazykových prostředků je též pro různé styly odlišná, viz (Těšitelová, 1987). Přesto ale právě fakt, že se jedná o báseň a navíc báseň svázanou jistými předem stanovenými kritérii, se jeví jako nevýhoda pro překladatele, ale výhoda pro ty, kteří se chtějí zabývat kvantitativní analýzou. Specifičnost výběru tohoto typu souboru do jisté míry potlačuje kritéria psychologická, sociologická, tematická i „sémiotická".

Pokud se jedná o rozhodování, zda je báseň *Havran* z hlediska statistického souborem základním či výběrovým, je jasné, že bychom si přáli, aby se kvantitativní výzkum týkal celého jazyka, tedy celého základního souboru, což je požadavek takřka neřešitelný. Kterýkoli z překladů či samotný originál stanovme výběrovými soubory. Co je základní soubor? Zde je důležitá úvaha a počáteční stanovení podmínek. Je nemyslitelné uvažovat o celém jazyce. Patrně též nemůžeme považovat za základní soubor všechny autorské texty Edgara Allana Poea, protože zkoumáme i překlady jiných autorů, byť respektující Poeova kritéria. Tedy stanovme si jako základní soubor text básně *The Raven* ve všech jazycích, do kterých byl přeložen.

V tomto experimentu se do zorného pole fraktální analýzy dostalo dvacet textů Poeova *Havrana* ve třech jazycích; jeden originální Poeův text v anglickém jazyce, překlad Otty F. Bablera do německého jazyka, viz (Poe, 1931), a osmnáct překladů do českého jazyka od různých autorů[24], viz (Poe, 1985), (Poe, 2008a) a (Poe, 2008b).

Přes všechno výše zmíněné, byla práce s poetickým textem velice obtížná, například stanovení jednotek a počítání délek některých entit, jak bude patrné dále. Proto byl zvolen pro porovnání text žurnalistický, který se nejen z hlediska sémantičnosti podstatně liší od textu poetického. Jedná se o náhodně zvolený článek z regionálního Svitavského deníku, viz (Nebeský, 2009).

---

považována za jeden z největších problémů translatologie. Už jen název *The Raven* je ve velké většině případů do českého jazyka překládán nesprávně jako „havran", jednou z výjimek je například překlad Miroslava Macka, (Poe, 1993), který použil slovo „krkavec". Důvodem pro hojné používání překladu „havran" může být česká tradiční literární symbolika, kdy havran je symbol a posel zla na rozdíl od krkavce. Dalším známým překladatelem, který v současnosti použil doslovný překlad „krkavec", je Tomáš Jacko, (Poe, 2008a). Jeho překlad byl navržen na cenu Josefa Jungmana.

[23] Klíčová slova básně, jako například „havran", jméno milenky, překlady refrénu „nevermore", a dostupnost jejich jednoznačných ekvivalentů v českém jazyce mě inspirovaly k tomu, abych text básně a speciálně tato slova či fráze podrobila též kvantitativnímu zkoumání z hlediska míry přenášené informace, viz kapitola 5.

[24] Jedním z překladatelů do českého jazyka je opět Otto F. Babler. Jeho překlad je z roku 1930, viz (Poe, 1985). Mimo zkoumání překladů jednoho textu do různých jazyků různými autory nám tudíž byla dopřána příležitost analyzovat překlad jednoho textu jedním autorem do dvou typově odlišných jazyků.

4.2.2 Krok 2 – stanovení jednotek

Jak bylo již psáno v části teoretické, stanovení jednotek je jeden z nejtěžších úkolů pro experimenty kvantitativní lingvistiky. V následujících odstavcích a v přílohách budu prezentovat a komentovat výstupy kvantifikace nahlížené čtyřmi přístupy. Tyto přístupy byly aplikovány na výše zmíněné výběrové soubory novinového článku, Poeova originálu *Havrana* a na devatenáct překladů Poeova *Havrana* do českého a německého jazyka. Data získaná kvantifikací těchto výběrových souborů jsou prezentována v tabulkách v příloze I.

Přístup 0. byl uplatněn na originální anglický text básně *Raven* a na žurnalistický text (Nebeský, 2009), data získaná kvantifikací obou výběrových souborů jsou k dispozici v tab. č. 19 a 20 v příloze I. Důvodem pro použití pravděpodobně z pohledu lingvistiky ne zcela nejefektivnější metody byla snaha získat počáteční data pro porovnání se „sofistikovanějšími" metodami a primárně na úvod experimentu prověřit algoritmus zpracování textů a stanovit jeho fundamentální součásti, aniž by se příliš akceptovalo definování jednotek, které by odpovídalo záměru experimentu. Na druhou stranu není vhodné zcela zamítnout tuto metodu před získáním většího množství dat, které by potvrdilo nebo vyvrátilo účelnost používání této metody. Tento přístup je uveden hlavně pro ilustraci a pro kontrast s ostatními přístupy, přijatelné výsledky však nepřinesl.

Přístup I. byl aplikovaný na originální text Poeovy básně *Raven*, na všech jeho šestnáct českých překladů dostupných v (Poe, 1985) a na jeho německou mutaci, (Poe, 1931). Všechny tabulky s výstupem shromážděných dat tímto způsobem jsou uvedené v příloze I. v tab. č. 21, …, 40.

Přístup III. byl aplikován na výše zmíněný žurnalistický text (Nebeský, 2009). V tomto vzorku, který byl nejprve podroben analýze založené na definování slova přístupem 0. (viz tab. č. 19_1, 19_2, 19_3 v příloze I.), jsme získali na úrovni slovo – slabika $b_3 < 0$ (viz tab. č. 19_3 a přehled parametrů b_i v kroku 5), to znamená, že by nešlo o jazykový fraktál, což vedlo k úvaze o správnosti postupu. V tomto vzorku se zhruba čtyřicet procent všech jednoslabičných slov sestávalo z předložek, které byly jedno nebo maximálně dvoufonémové. V celkovém počtu všech výskytů počet slabik nově vytvořených tvarů složených z předložek a následujícího slova z textu nepřesáhl jejich počet v původních slovech, která nyní přijala předcházející předložku, neboť předložky byly z velké části neslabičné (jako v případě *v čem*). Nově vzniklé složeniny musí být považovány za slovní jednotky, abychom předešli ztrátě jakéhokoli fonému. Výstup získaný tím to způsobem je ilustrován v příloze I. tab. č. v 43_1, 43_2, 43_3, viz také (Andres et al., 2011). Přístup III. byl dále aplikován na originální Poeův text *Havrana*, viz příloha I. tab. č. 44_1, 44_2, 44_3, na německý, tab. č. 45_1, 45_2, 45_3, a český Bablerův překlad, tab. č. 46_1, 46_2, 46_3. Protože se, jak bude vidět dále, ukázal být velice efektivním, je navrženo pro další budoucí experimenty kvantifikovat i ostatní české mutace a další výběrové soubory s použitím tohoto přístupu.

Přístup II. byl testován na stejných vzorcích jako přístup III., s výjimkou žurnalistického textu a Bablerova překladu *Havrana* do českého jazyka z důvodu absence členů v českém jazyce. Výstupy tohoto přístupu jsou publikovány v příloze I. v tab. č. 41 a 42.

V tabulkách, viz příloha I. tab. č. 47, přináším pro porovnání výstupy kvantifikace originálního textu Poeova *The Raven* s přihlédnutím k diskusi o stanovení jednotek syntaktické úrovně. Věta je zde definována jako nejmenší možný segment daného výběrového souboru, jehož řídící člen připouštíme mimo finitního slovesného tvaru i tvar infinitní.

Jeden z nejobtížnějších kroků celého algoritmu je kvantifikace sémantických konstruktů, proto považuji za vhodné nastínit její fundamenty. Postup kvantifikace binarismu sémantické konstrukty (ve větách/klauzích) – věty/klauze (měřené v průměrném počtu jejich slov)[25]:

1. Každému slovu přiřadíme číslo, které udává počet slov ve větě, kde se dané slovo aktuálně vyskytuje.
2. Transformujeme slova do jejich základní podoby většinou shodné s podobou slovníkovou, tj. do tvaru jejich lexému. Tento proces se nazývá lemmatizace.
3. Spočítáme, kolikrát se každý lexém vyskytuje. Toto číslo se nazývá frekvence/četnost výskytu daného lexému, viz tab. č. 3. Tato čísla jsou veličiny $x_{1j}, j \in N$. Jinými slovy, například $x_{11} = 1$ znamená jednovětý sémantický konstrukt, tedy sémantický konstrukt, který se skládá právě z jedné věty, tedy sémantický konstrukt, který je postaven na lexému vyskytujícím se právě jednou (právě a pouze v jediné větě). Příslušné z_{11} je počet všech jednovětých sémantických konstruktů.

lexémy	četnost každého lexému	počet slov v klauzích daného lexému
at	8	114
he	8	63
Lenore	8	75
much	8	60
or	8	100
then	8	62
with	8	144
bird	10	169
on	10	110
raven	10	80
chamber	11	195
nevermore	11	86
be to	14	105
door	14	206
a	15	193
that	17	177
this	18	183
of	20	298
my	24	284
and	38	376

[25] Ke kvantifikaci výběrového souboru je doporučen například funkce Microsoft Excel COUNTIF, SUMIF.

I	41	294
the	56	566

Tab. č. 3: E.A. Poe – ukázka části tabulky pro kvantifikaci binarismu sémantický konstrukt – věta/klauze pro nejčetnější lexémy pro $x_{18} = 8, x_{19} = 10, x_{110} = 11, x_{111} = 14, x_{112} = 15, x_{113} = 17, x_{114} = 18, x_{115} = 20, x_{116} = 24, x_{117} = 38, x_{118} = 41, x_{119} = 56$

4. Pro každé x_{1j} spočítáme průměrnou délku příslušných vět ve slovech, tj. y_{1j}. Například pro $x_{11} = 1$ spočteme průměr čísel, které byly přiřazeny v bodě 1. všem lexémům vyskytujícím se v textu právě jednou, a dostaneme tak y_{11}.
5. Výsledná tabulka tedy obsahuje x_{1j} a příslušné z_{1j} a y_{1j} připravené pro další vyhodnocení, viz tab. č. 4.

j	x_{1j}	z_{1j}	y_{1j}
1	1	250	11,168
2	2	72	11,04167
3	3	27	10,23457
4	4	13	11
5	5	11	10,85455
6	6	6	11,5
7	7	11	10,5974
8	**8**	**7**	**11,03571**
9	**10**	**3**	**11,96667**
10	**11**	**2**	**12,77273**
11	**14**	**2**	**11,10714**
12	**15**	**1**	**12,86667**
13	**17**	**1**	**10,41176**
14	**18**	**1**	**10,16667**
15	**20**	**1**	**14,9**
16	**24**	**1**	**11,83333**
17	**38**	**1**	**9,894737**
18	**41**	**1**	**7,170732**
19	**56**	**1**	**10,10714**

Tab. č. 4: E.A. Poe – výsledná tabulka binarismu sémantický konstrukt – věta/klauze (tučně řádky odpovídající tab. č. 3)

V další pasáži je nutné se podrobněji zmínit o procesu a alespoň některých pravidlech lemmatizace uplatněných v tomto experimentu. Samotný proces se může lišit dle typu jazyka. Je třeba mít na zřeteli, že náš experiment by měl odrážet sémantickou hustotu výběrového

a potažmo i základního souboru. Uvádíme několik pravidel a typických příkladů lemmatizace výběrového souboru v českém i anglickém jazyce, pokud se liší[26].

- V českém jazyce se substantiva, zájmena a číslovky uvádějí v nominálu singuláru, (Těšitelová, 1987, s.13).
- V českém jazyce se adjektiva, zájmena a číslovky adjektivní povahy uvádějí v nominálu singuláru maskulina, (Těšitelová, 1987, s.13).
- Komparativ a superlativ adjektiv a adverbií se uvádějí jako pozitiv (*dál → daleko)*.[27]
- V českém jazyce se slovesné tvary, včetně transgresív (*hledaje → hledat*), participií, pasiv (*je zastřeno → zastřít*) uvádějí ve tvaru infinitivu, (Těšitelová, 1987, s.13), (Petr et al., 1986b, s. 416-427).
- Adjektiva tvořená z přechodníků jsou ponechána jako adjektiva, (*dorozumívající se*), (Petr et al., 1986b, s. 416-427).
- Substantiva verbale (podst.jm.slovesná) jsou ponechána jako substantiva, (Petr et al., 1986b, s. 416-427).
- Adverbia utvořená od adjektiv jsou chápána jako samostatná slova.
- Deminutiva, augmentativa a přechýlená substantiva jsou uváděna zvlášť.

- *abych, abys, … → aby*
- *ve → v, se → s,* atd.
- Spojkový výraz *-li* je uváděn zvlášť.

- Pro anglický jazyk uvádí Těšitelová, že díky chudému tvarosloví se zvlášť mohou uvádět i plurály substantiv, pravidelně tvořené tvary komparativů a superlativů adjektiv a adverbií, adverbia odvozená koncovkou *–ly*, viz (Těšitelová, 1987, s.13). Pro náš experiment je doporučeno řešit tyto problémy stejným způsobem jako u českého výběrového souboru, aby bylo zachováno společné sémantické pozadí výběrů.
- Speciální, velice důležitý problém pro výběrové soubory v anglickém jazyce je problém kvantifikace členů, jak již bylo zmíněno výše. Člen v anglickém jazyce plní sémanticko-gramatickou funkci a „syntakticky má funkci determinátoru, tj. nesamostatného větného členu v rámci větného členu realizovaného substantivem, v němž zpravidla tvoří první složku, tj. předchází před premodifikací“, viz (Dušková, 1994). Z toho vyplývá, že by též přicházelo v úvahu počítat jako jeden znak člen dohromady se substantivem, které rozvíjí, jak také navrhuje Hřebíček[28], viz (Hřebíček, 1997). Dle Těšitelové, (Těšitelová, 1987, s.15), se ale pokládá většinou za samostatnou jednotku, k čemuž se prozatím ve většině případů v našem experimentu přikloníme, aby pak nebylo nutné například fráze typu *a raven* a *the raven* počítat zvlášť. Druhým důvodem je snaha o konzistentnost s vyšetřováním výběru pomocí teorie informace, viz dále.

[26] V anglickém jazyce se ve velké míře stírá rozdíl mezi slovoformami a příslušnými lexémy.

[27] Dle Marie Těšitelové je ale možné též při lemmatizaci pozitiv uvádět jako jeden tvar a komparativ a superlativ uvádět jako tvar druhý, nebo všechny tři tvary uvádět zvlášť, (Těšitelová, 1987, s.14).

[28] Dle Duškové, viz (Dušková, 1994), se člen dále od premodifikátoru liší tím, že příslušný větný člen v angličtině provází obligatorně. Pokud tedy budeme počítat se členy jako znaky nebo částmi znaků, je třeba zvážit nulovou variantu členu neurčitého a její započítání jako znak či součást znaku.

4.2.3 Krok 3 – test reprezentativnosti výběrového souboru

Postup budiž ilustrován na originálním textu E. A. Poea *The Raven*, ve kterém slova nechť jsou definována jako jednotlivé slovoformy, tedy analyzovaném pomocí přístupu I. Každé slovoformě opět přiřadíme příslušný odpovídající lexém a zjistíme pravděpodobnost jejího výskytu ($p_i = \frac{N_i}{N}$, kde N_i je četnost jednotlivých lexémů a N je celkový počet lexémů), viz tab. č. 5.

lexémy	N_i	p_i
at	8	0,007547
he	8	0,007547
Lenore	8	0,007547
much	8	0,007547
or	8	0,007547
then	8	0,007547
with	8	0,007547
bird	10	0,009434
on	10	0,009434
raven	10	0,009434
chamber	11	0,010377
nevermore	11	0,010377
be to	14	0,013208
door	14	0,013208
a	15	0,014151
that	17	0,016038
this	18	0,016981
of	20	0,018868
my	24	0,022642
and	38	0,035849
I	41	0,038679
the	56	0,05283

Tab. č. 5: E.A. Poe – četnosti a pravděpodobnosti nejfrekventovanějších lexémů

Do vzorce (29) dosadíme $k = 412$ a pravděpodobnosti jednotlivých lexémů (viz příloha X.), tzn., $\sum_{i=1}^{412} \ln p_i = -2665{,}03$. Stanovme si $r = 0{,}0012$. Tedy

$$\ln N = \frac{1}{412 - 1} \cdot (-2665{,}03) - 2 \cdot \ln 0{,}0012$$

$$\ln N = 6{,}96661$$

$$N = e^{6{,}96661}$$

$$N = 1060{,}621$$

Náš výše zmíněný výběrový soubor obsahuje celkově 1 060 lexémů. Takový výběrový soubor je tedy při předem stanovené průměrné směrodatné odchylce $r = 0{,}0012$, tj. při průměrné směrodatné odchylce 0,12%, stabilní a reprezentativní.

Problém výběru reprezentativního vzorku je jedním z nejdůležitějších, avšak v naší analýze byla prvotní motivací pro výběr vzorků jedinečná šance analyzovat různé texty v různých jazycích s totožným sémantickým pozadím.

4.2.4 Krok 4 – kvantifikace výběrových souborů

Na základě stanovení jednotek výše popsanými přístupy jsou kvantifikovány texty a získány tabulky odrážející výsledky na rovinách všech třech zmíněných binarismů obsahující konstrukty s délkami x_i, jejich frekvence z_i a konstituenty s délkami y_i, pro každý binarismus označený $i = 1{,}2{,}3$, viz tabulky v příloze I. výsledků kvantifikace originálního textu E.A. Poea *The Raven*, překladů do českého jazyka, Bablerova překladu do německého jazyka a žurnalistického textu (Nebeský, 2009). Krok 5 – výpočet parametrů A_i, b_i, c_i, pro i = 1, 2, 3

4.2.4.1 Výpočet pomocí statistických metod

Pro názornost demonstruji celý výpočet na příkladu výběrového souboru originálního textu E. A. Poea *The Raven*. Zvolme si například binarismus $i = 1$, sémantický konstrukt – věta/klauze. Tab. č. 6 je rozšířenou variantou tab. č. 4. Je rozšířena o mezivýpočty potřebné pro dosazení do vzorců pro výpočet koeficientů a, b.

j	x_j	$x'_j = \ln x_j$	$x_j{}'^2$	z_j	y_j	$y'_j = \ln y_j$	$x'_j y'_j$
1	1	0	0	250	11,168	2,413053	0
2	2	0,693147	0,480453	72	11,04167	2,401676	1,664715
3	3	1,098612	1,206949	27	10,23457	2,325771	2,555121
4	4	1,386294	1,921812	13	11	2,397895	3,324189
5	5	1,609438	2,59029	11	10,85455	2,384584	3,83784
6	6	1,791759	3,210402	6	11,5	2,442347	4,376098
7	7	1,94591	3,786566	11	10,5974	2,360609	4,593533
8	8	2,079442	4,324077	7	11,03571	2,401137	4,993024
9	10	2,302585	5,301898	3	11,96667	2,482125	5,715304
10	11	2,397895	5,749902	2	12,77273	2,547312	6,108188
11	14	2,639057	6,964624	2	11,10714	2,407588	6,353764
12	15	2,70805	7,333536	1	12,86667	2,55464	6,918093
13	17	2,833213	8,027098	1	10,41176	2,342936	6,638039
14	18	2,890372	8,354249	1	10,16667	2,319114	6,703103
15	20	2,995732	8,974412	1	14,9	2,701361	8,092555
16	24	3,178054	10,10003	1	11,83333	2,47092	7,852718
17	38	3,637586	13,23203	1	9,894737	2,292003	8,337358
18	41	3,713572	13,79062	1	7,170732	1,970008	7,315766
19	56	4,025352	16,20346	1	10,10714	2,313242	9,311614
Σ		43,92607	121,5524			45,52832	104,691
kvadrát		1929,5					

Tab. č. 6: E.A. Poe – rozšířená tabulka výsledků kvantifikace textu zvoleným způsobem

Z tabulky snadno přečteme:

$$n = 19$$

$$\sum_{j=1}^{19} x'_j = 43{,}92607$$

$$\sum_{j=1}^{19} y'_j = 45{,}52832$$

$$\sum_{j=1}^{19} {x'_j}^2 = 121{,}5524$$

$$\left(\sum_{j=1}^{19} x'_j\right)^2 = 1929{,}5$$

$$\sum_{j=1}^{19} x'_j y'_j = 104{,}691$$

Dosazením do vzorců (40) a (41) dostáváme

$$b = -\frac{19 \cdot 104{,}691 - 43{,}92607 \cdot 45{,}52832}{19 \cdot 121{,}5524 - 1929{,}5} = 0{,}02829,$$

$$a = \frac{121{,}5524 \cdot 45{,}52832 - 104{,}691 \cdot 43{,}92607}{19 \cdot 121{,}5524 - 1929{,}5} = 2{,}461637.$$

Jelikož $a = \ln A'$, obráceným postupem dostaneme

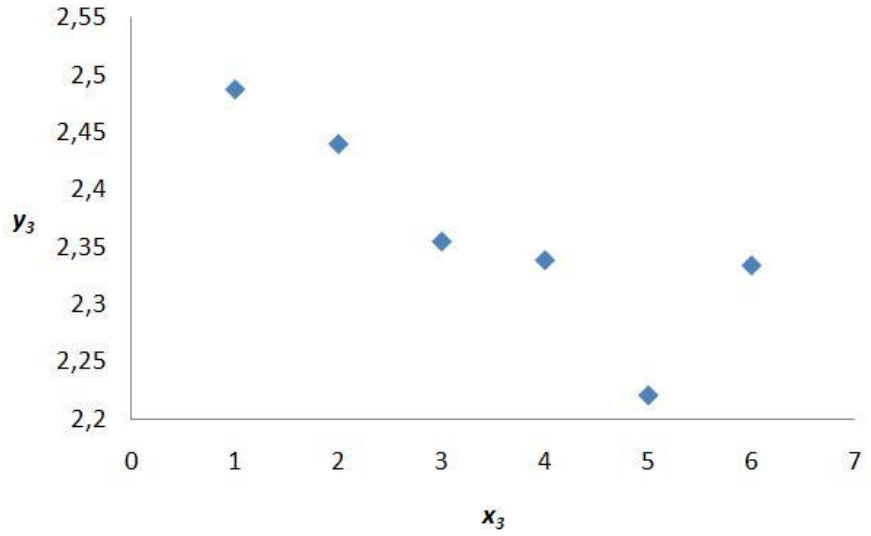
$$A' = e^{2{,}461637} = 11{,}72398.$$

Výše uvedené výsledky byly získány pomocí Microsoft Excelu, samozřejmě je možno též použít kalkulačku. Byl by to ale velice zdlouhavý a pracný postup. Namísto toho se velice elegantně a efektivně dá použít statického softwaru, například R (ten byl použit v našem experimentu) nebo SAS. V příloze II. je k nahlédnutí program sloužící pro získání koeficientů A', b z tabulek výstupů kvantifikace různých textů[29]. Používaný software byl R 2.10.0, který je volně dostupný na internetu, je možné jej stáhnout na www.r-project.org.

V příloze III. je k dispozici ukázka výstupu statistického softwaru R 2.10.0. Pro velkou rozsáhlost získaných dat byl vybrán jediný výstup, který dostatečně ilustruje získaná data a je příslušný k výběrovému souboru Poe 1a, b, c, d I., tedy výběrovému souboru nahlíženému

[29] Tento program se vlastně skládá ze šesti částí, jednoduchá verze MAL – lineární regrese, logaritmizace, jednoduchá verze MAL – nelineární regrese, jednoduchá verze MAL – Taylorův rozvoj, úplná verze MAL – lineární regrese, logaritmizace, úplná verze MAL – nelineární regrese, úplná verze MAL – Taylorův rozvoj. Podrobně je diskutována obzvláště první metoda, která je v případě nutnosti nejvhodnější pro ruční výpočty.

prostřednictvím přístupu I. a zpracovanému metodami statistickými i numerickými skrze jednoduchou i úplnou verzi MAL. Červeně jsou zvýrazněny vyčíslené parametry, které jsou doplněny pro úplnost na závěr každé metody. Zde jsou též pro ilustraci zařazeny metody znázorňující hledané regresní křivky. I grafy byly získány pomocí software R. Ostatní číselné výstupy programu R budou komentovány v další kapitole.



Obr. č. 7: Izolované body odpovídající pozorováním z tab.č. 43_3

Velice krátce se seznamme s fungováním tohoto softwaru a motivací při jeho tvoření. Jako příklad volím analýzu dat z tab. č. 43_3 (vztah slova – slabiky). Izolované body znázorňující naše pozorování jsou vyneseny v grafu na obr. č. 7. Jednoduchý způsob, kterým je možné zadat do softwaru data, je nechat je načíst z jednoduchého textového souboru s koncovkou .txt. Předpokládejme, že soubor obsahuje dva sloupce (což je vlastně tabulka obsahující každou hodnotu našich pozorování), kde první sloupec odpovídá délce slov ve slabikách (proměnná `x`) a druhý sloupec obsahuje délky slabik ve fonémech (proměnná `length`), každý řádek odpovídá jednomu slovu v analyzovaném výběrovém souboru, jako např.

```
"x" "length"
1   1
2   3
3   5
...,
```

kde první řádek je záhlaví obsahující názvy proměnných. Tento soubor (pojmenovaný "`text_2_3.txt`") může být do software R načten příkazem

```
text=read.table("text_2_3.txt",header=T,sep="\t").
```

Nejprve je třeba vypočítat poměr length/*x* jako nová proměnná `y` v `text` datovém rámci příkazem `text=cbind(text,y=text$length/text$x)`. Datový rámec `tabY` odpovídající hodnotám z tabulky 43_3

```
x      avg

1 2.486957

2 2.439227

3 2.354167

4 2.337963

5 2.220000

6 2.333333
```

je poté vytvořen následujícím kódem

```
> x=as.numeric(levels(as.factor(text$x)))

> avg=as.numeric(tapply(text$y,text$x,FUN=mean))

> tabY=data.frame(x=x,avg=avg).
```

Nyní jednoduše nalezneme odpovídající lineární model pomocí fukce `lm()`

```
> model1=lm(log(tabY$avg) ~ log(tabY$x))

> model2=lm(log(tabY$avg) ~ log(tabY$x)+tabY$x)
```

a získáme odpovídající odhadnuté hodnoty ***β*** funkcí `coef()`

```
> coef(model1)

(Intercept) log(tabY$x)

 0.91515690 -0.05136281

> coef(model2)

 (Intercept)  log(tabY$x)       tabY$x

 0.913375549 -0.061009841  0.003531349,
```

což znamená, jak již bylo výše ukázáno na jiném případu a jiném výběrovém souboru, že výsledky (hodnoty parametrů modelu odhadnuté pomocí metody nejmenších čtverců) pro zkrácený tvar formule MAL jsou: $\ln(A_3)$ = 0.91515690..., b_3 = 0.05136281..., a pro úplný tvar formule MAL jsou: $\ln(A_3)$ = 0.913375549..., b_3 = 0.061009841..., c_3 = 0.003531349....

4.2.4.2 Výpočet numerickými metodami

Pro komplikovanost numerických metod bude výpočet demonstrován výhradně za pomoci statistického softwaru R. To, aby náš model dobře aproximoval datové soubory `text`, může být spolehlivě zajištěno funkcí `nls()`, která poskytuje Gauss-Newtonův algoritmus a který umožňuje vyřešit nelineární problém nejmenších čtverců, viz (Stoer & Bulirsch, 2002). Podržme si naposledy použitý výběrový soubor, pro jednoduchou verzi formule MAL sestavme následující sekvenci příkazů

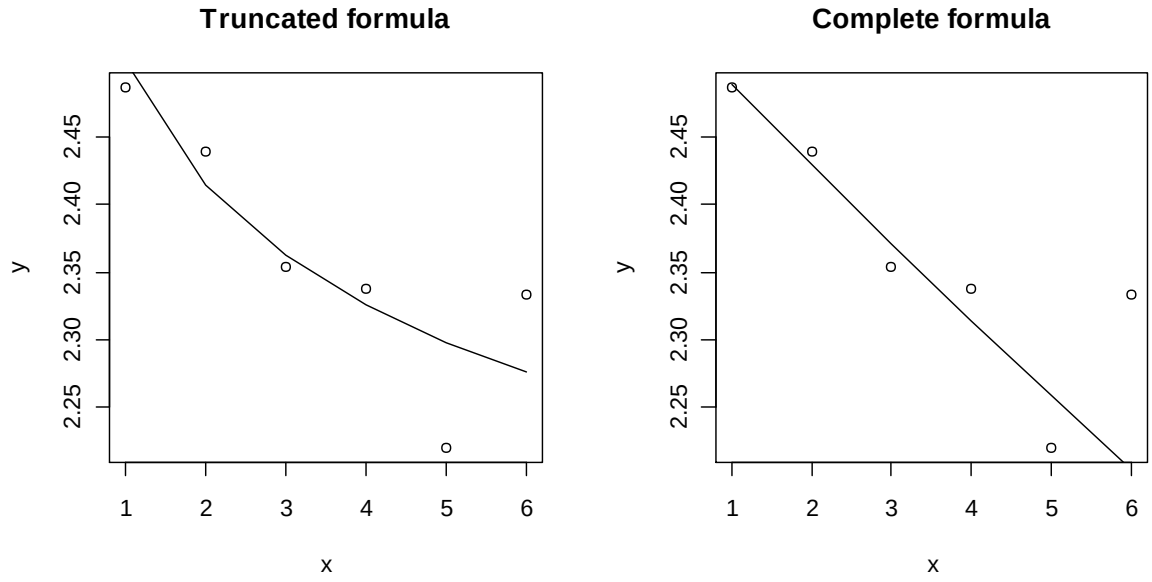
```
> model1.nls=nls(y ~ A*x^(-b), data=text,
start=list(A=exp(coef(model1)[1]),b=-coef(model1)[2]))

> summary(model1.nls)$coefficients[,1]
```

```
          A           b

2.50621226 0.05390454
```

a pro úplnou verzi formule MAL

```
> model2.nls=nls(y ~ A*x^(-b)*exp(c*x), data=text,
start=list(A=exp(coef(model2)[1]),b=-coef(model2)[2],

c=coef(model2)[3]))

> summary(model2.nls)$coefficients[,1]

             A              b              c

 2.5516707542 -0.0004874368 -0.0245950992.
```

Získané křivky jsou demonstrovány na obr. č. 8.



Obr. č. 8: Grafické porovnání modelů jednoduché a úplné verze formule MAL – Gauss-Newtonův algoritmus

4.2.5 Přehled a komentáře k vypočteným hodnotám parametrů

Celkové výsledky kvantifikace originálního textu E. A. Poea *The Raven* zpracované danými statistickými metodami při daném stanovení jednotek dle přístupu I. jsou publikovány v tab. č. 20 v příloze I. Abychom testovali možnou souvislost jazykových a fraktálních struktur, prozkoumáme matematický fraktál, který je za určitých podmínek přidružený dané jazykové struktuře, poté, co jsme vypočítali jeho fraktální dimenzi. Znovu zdůrazňujeme, že model zkoumané jazykové struktury může pouze aproximovat s dostatečnou přesností tento matematický fraktál, neboť u jazykové struktury v našem experimentu zkoumáme pouze první iteraci (výše definované tři binarismy), které jsme dosud měli k dispozici. Nevylučujeme

a vítáme možné rozšíření aparátu v budoucích experimentech, viz závěr. Jak již bylo zmíněno výše, jazykovým fraktálem je takový lingvistický subjekt, který splňuje MAL se všemi b_i na všech svých úrovních kladnými, viz (Andres, 2009). Dále pak soběpodobnostní dimenze *D* budiž mírou sémantičnosti textu, kterou je možné vypočítat pomocí formule (26) jako reciprokou hodnotu aritmetického průměru koeficientů b_i, viz (Andres, 2009).

V pojetí Luďka Hřebíčka, viz (Hřebíček, 1997), (Hřebíček, 2002), je každému binarismu přiřazena dimenze (reciproká hodnota příslušného koeficientu b_i, která opět musí být kladná), aby tedy struktura byla tímto prohlášena fraktálem. Takovýto fraktál je lépe nazvat slabou variantou fraktálu, neboť sice text splňuje na všech svých jazykových hladinách MAL, ale na každé úrovni má jinou fraktální dimenzi $D_i,\ i = 1{,}2{,}3$, viz (Andres, 2010). Přesto však při hodnocení výsledků našeho experimentu považujeme za nutné připojit i čísla označená jako $D_i, i = 1{,}2{,}3$, abychom pomocí těchto čísel mohli sledovat, jak mnoho kolísá sémantičnost na všech úrovních v rámci jednoho výběrového souboru a jednoho způsobu stanovení jednotek. Tato čísla však nebudeme nazývat dimenzemi jazykového fraktálu.

Jazykový fraktál ve své silné variantě splňuje MAL a zároveň má $b_i > 0$, pro všechna $i = 1{,}2{,}3$, jak již několikrát bylo zmíněno výše. Vizualizovaný model tohoto fraktálu je aproximací přidruženého matematického fraktálu a sémantičnost způsobuje, že jeho *D* roste. Je to ovšem dimenze matematického fraktálu, která je jazykovému fraktálu pouze přiřazena (jehož je vizualizovaný model jazykového fraktálu aproximací) a reflektuje bohatost struktury z hlediska sémantiky. Můžeme tedy říci, že text je sémanticky tak bohatý, jak vysoká je dimenze přidruženého matematického fraktálu, viz (Andres, 2010).

Podívejme se tedy, jak to vypadá s fraktálností a sémantičností originálního textu E. A. Poea *The Raven* s jednotkami stanovenými dle přístupu I.

Jednoduchá verze MAL - logaritmizace

POEI		A_i	b_i	c_i	D_i
sémantické konstrukty - klauze	jednoduchá verze MAZ - logaritmizace	11,7240	0,0283		35,3452
klauze - slova	jednoduchá verze MAZ - logaritmizace	1,6609	0,0491		20,3577
slova - slabiky	jednoduchá verze MAZ - logaritmizace	2,6179	0,0685		14,5942

Tab.č. 7$_a$: E.A. Poe (přístup I.) – výsledné hodnoty koeficientů A_i, b_i jednoduché verze MAL metodou lineární regrese

V případě výsledků z tab. č. 7$_a$ je možné říci, že zkoumaný výběrový vzorek je jazykovým fraktálem, neboť splňuje MAL a všechny koeficienty $b_i, i = 1{,}2{,}3$ jsou kladné. Fraktální dimenze, která je textu přiřazená je

$$D = \frac{3}{0{,}0685204+0{,}04912137+0{,}02829237} = 20{,}55722.$$

Jednoduchá verze MAL - numerické řešení MNČ

POEI		A_i	b_i	c_i	D_i
sémantické konstrukty - klauze	jednoduchá verze MAZ - numerické řešení MNČ	11,5713	0,0186		53,7346
klauze - slova	jednoduchá verze MAZ - numerické řešení MNČ	1,8095	0,0845		11,8301
slova - slabiky	jednoduchá verze MAZ - numerické řešení MNČ	2,6184	0,0680		14,7124

Tab.č. 7_b: E.A. Poe (přístup I.) – výsledné hodnoty koeficientů A_i, b_i jednoduché verze MAL metodou nelineární regrese

V případě výsledků z tab. č. 7_b je možné říci, že zkoumaný výběrový vzorek je jazykovým fraktálem, neboť splňuje MAL a všechny koeficienty $b_i, i = 1,2,3$ jsou kladné. Fraktální dimenze, která je textu přiřazená, je

$$D = \frac{3}{0{,}06797 + 0{,}08453 + 0{,}01861} = 17{,}53258.$$

Úplná verze MAL - logaritmizace

POEI		A_i	b_i	c_i	D_i
sémantické konstrukty - klauze	úplná verze MAZ - logaritmizace	10,4692	-0,0824	0,0090	-12,1376
klauze - slova	úplná verze MAZ - logaritmizace	1,8430	0,1950	-0,0186	5,1278
slova - slabiky	úplná verze MAZ - logaritmizace	2,5808	0,1048	-0,0172	9,5435

Tab.č. 7_c: E.A. Poe (přístup I.) – výsledné hodnoty koeficientů A_i, b_i, c_i úplné verze MAL metodou lineární regrese

V případě výsledků z tab. č. 7_c je možné říci, že zkoumaný výběrový vzorek není jazykovým fraktálem, neboť sice splňuje MAL, ale koeficient b_1 je záporný. Fraktální dimenzi tedy nemá smysl počítat.

Úplná verze MAL - numerické řešení MNČ

POEI		A_i	b_i	c_i	D_i
sémantické konstrukty - klauze	úplná verze MAZ - numerické řešení MNČ	10,3038	-0,0951	0,0095	-10,5106
klauze - slova	úplná verze MAZ - numerické řešení MNČ	2,0419	0,2672	-0,0238	3,7422
slova - slabiky	úplná verze MAZ - numerické řešení MNČ	2,5952	0,0897	-0,0105	11,1520

Tab.č. 7_d: E.A. Poe (přístup I.) – výsledné hodnoty koeficientů A_i, b_i, c_i úplné verze MAL metodou nelineární regrese

V případě výsledků z tab. č. 7_d je možné říci, že zkoumaný výběrový vzorek není jazykovým fraktálem, neboť sice splňuje MAL, ale koeficient b_1 je záporný. Fraktální dimenzi tedy nemá smysl počítat.

Výsledky kvantifikace ostatních textů jsou zveřejněny v příloze IV. Ve výše zmíněném případě originálního Poeova textu můžeme s ohledem na sémantičnost textu porovnávat jen vyhodnocení pomocí jednoduché verze MAL oběma preferovanými způsoby, tj. lineární a nelineární regresí. Porovnání fraktálních dimenzí všech zkoumaných výběrových souborů (tam, kde to má smysl) bude též zveřejněno dále.

Poté, co byly všechny výběrové soubory kvantifikovány tak, jak bylo naznačeno v kroku 4, byl výstup zpracován čtyřmi způsoby.

a Výpočet parametrů A_i, b_i, i = 1, 2, 3 pro jednoduchou verzi formule MAL pomocí statistických metod (v grafech vyznačeno jako ◯).
b Výpočet parametrů A_i, b_i, i = 1, 2, 3 pro jednoduchou verzi formule MAL pomocí numerických metod (v grafech vyznačeno jako □).
c Výpočet parametrů A_i, b_i, c_i, i = 1, 2, 3 pro úplnou verzi formule MAL pomocí statistických metod (v grafech vyznačeno jako ◇).
d Výpočet parametrů A_i, b_i, c_i, i = 1, 2, 3 pro úplnou verzi formule MAL pomocí numerických metod (v grafech vyznačeno jako △).

Jak již bylo několikrát zmíněno, nejdůležitějším z parametrů je b_i, pro i = 1, 2, 3 z důvodů své korelace s dimenzí přidruženého matematického fraktálu. Z tohoto důvodu je v následujícím odstavci prezentován jen tento parametr[30] pro všechny výběrové soubory zpracované jedním z dříve zmíněných přístupů[31].

Ad přístup I.:

1 Poe
   a. b_1= 0,02829237, b_2= 0,04912137, b_3= 0,0685204
   b. b_1= 0,01861, b_2= 0,08453, b_3= 0,06797
   c. (b_1= *-0 ,08238833*, b_2= 0,1950147, b_3= 0,1047829)
   d. (b_1= *-0, 095142*, b_2= 0,26722, b_3= 0,08967)
2 Babler – německý
   a. (b_1= 0,04469562, b_2= *-0 ,003084549*, b_3= 0,2741698)
   b. (b_1= 0,0353, b_2= *-0,002524*, b_3= 0,2839)
   c. b_1= 0, 1321585, b_2= 0, 1568826, b_3= 0, 3793558
   d. b_1= 0, 111254, b_2= 0, 1728, b_3= 0, 39072
3 Šembera
   a. b_1= 0,0185512, b_2= 0,04661435, b_3= 0,2434756
   b. b_1= 0,0102, b_2= 0,04521, b_3= 0,2411
   c. (b_1= *-0 ,0326087*, b_2= 0,0758137, b_3= 0,183899)

[30] Přehled všech parametrů pro všechny výběrové soubory viz přílohy IV.
[31] V závorkách jsou prezentovány výstupy kvantifikace, kde přinejmenším jeden z parametrů b_i, i = 1, 2, 3 je záporný (tyto jsou zvýrazněny kurzívou). Takovéto výběrové soubory nemohou tudíž být považovány za jazykové fraktály.

d. (b_1= *-0 ,044947*, b_2= 0,077833, b_3= 0,1957)

4 Vrchlický
   a. (b_1= 0,02785711, b_2= *-0 ,01016194*, b_3= 0,1988444)
   b. (b_1= 0,02811, b_2= *-0 ,01119*, b_3= 0,1966)
   c. b_1= 0,002215368, b_2= 0,05174956, b_3= 0,110384
   d. b_1= 0,002748, b_2= 0,05218, b_3= 0,12683

5 Mužík
   a. (b_1= *-0 ,002734985*, b_2= 0,06864244, b_3= 0,1325155)
   b. (b_1= *-0 ,02487*, b_2= 0,07662, b_3= 0,1254)
   c. (b_1= 0,09891378, b_2= 0,15536, b_3= *-0 ,1248322*)
   d. (b_1= 0,09682, b_2= 0,16274, b_3= *-0 ,1235*)

6 Lutinov
   a. b_1= 0,09267924, b_2= 0,02222886, b_3= 0,1504526
   b. b_1= 0,06622, b_2= 0,02643, b_3= 0,1462
   c. (b_1= 0,08755964, b_2= 0,2005795, b_3= *-0 ,03112276*)
   d. (b_1= 0,02428, b_2= 0,20572, b_3= *-0 ,03559*)

7 Nezval
   a. (b_1= 0,1772, b_2= *-0 ,02306*, b_3= 0,12116)
   b. (b_1= 0,1157, b_2= -0 ,0252, b_3= 0,1036)
   c. (b_1= 0,239787, b_2= 0,128708, b_3= *-0 ,52553*)
   d. (b_1= 0,175659, b_2= 0,12008, b_3= *-0 ,4916*)

8 Babler - český
   a. (b_1= *-0,01229989*, b_2= 0,07013482, b_3= 0,3309882)
   b. (b_1= *-0,03027*, b_2= 0,08325, b_3= 0,3049)
   c. (b_1= 0,3905951, b_2= 0,228655, b_3= *-0,2213671*)
   d. (b_1= 0,33339, b_2= 0,26447, b_3= *-0,1153*)

9 Taufer
   a. (b_1= 0,1610241, b_2= *-0 ,00942236*, b_3= 0,1290018)
   b. (b_1= 0,1058, b_2= *-0,009985*, b_3= 0,1238)
   c. b_1= 0,1693824, b_2= 0,01130945, b_3= 0,01274076
   d. (b_1= 0,06152, b_2= 0,006658, b_3 *cannot be found*)

10 Stoklas
   a. b_1= 0,1013934, b_2= 0,05913767, b_3= 0,0733
   b. b_1= 0,07163, b_2= 0,06786, b_3= 0,07232
   c. b_1= 0,1897575, b_2= 0,188745, b_3= 0,08140624
   d. b_1= 0,17795, b_2= 0,20283, b_3= 0,075108

11 Wagnerová
   a. (b_1= *-0 ,01852006*, b_2= 0,1014816, b_3= 0,08375543)
   b. (b_1= *-0 ,02034*, b_2= 0,1131, b_3= 0,08221)
   c. b_1= 0,005319176, b_2= 0,260446, b_3= 0,06177219
   d. (b_1= *-0 ,0001498*, b_2= 0,26399, b_3= 0,0476)

12 Havel
   a. b_1= 0,09848818, b_2= 0,05905367, b_3= 0,2610285

b. b_1= 0,05242, b_2= 0,06048, b_3= 0,2476
c. b_1= 0,1083344, b_2= 0,1159164, b_3= 0,09005535
d. b_1= 0,105941, b_2= 0,112434, b_3= 0,07996

13 Čapek
a. b_1= 0,0623626, b_2= 0,0330582, b_3= 0,07078767
b. b_1= 0,04679, b_2= 0,03114, b_3= 0,069
c. (b_1= 0,001749692, b_2= 0,1197062, b_3= *-0 ,01621862*)
d. (b_1= 0,001656, b_2= 0,11379, b_3= *-0 ,02259*)

14 Resler
a. b_1= 0,09714409, b_2= 0,02624885, b_3= 0,0868665
b. b_1= 0,06778, b_2= 0,02856, b_3= 0,084
c. (b_1= 0,2190468, b_2= 0,1446877, b_3= *-0 ,05152063*)
d. (b_1= 0,1521, b_2= 0,14568, b_3= *-0 ,06303*)

15 Černý
a. b_1= 0,04618615, b_2= 0,08846803, b_3= 0,005346203
b. (b_1= 0,04227, b_2= 0,104, b_3= *-0 ,0006068*)
c. b_1= 0,04812482, b_2= 0,2010607, b_3= 0,2348718
d. b_1= 0,028752, b_2= 0,2441, b_3= 0,25199

16 Slavík
a. b_1= 0,09306961, b_2= 0,09046977, b_3= 0,027177
b. b_1= 0,08111, b_2= 0,1345, b_3= 0,02541
c. (b_1= 0,1558316, b_2= 0,2727789, b_3= *-0 ,3087926*)
d. (b_1= 0,092506, b_2= 0,36934, b_3= *-0 ,3169*)

17 Kadlec
a. (b_1= *-0 ,04349735*, b_2= 0,08320275, b_3= 0,1611956)
b. (b_1= *-0 ,05865*, b_2= 0,09526, b_3= 0,1519)
c. (b_1= *-0 ,06038785*, b_2= 0,1269568, b_3= *-0,01063657*)
d. (b_1= *-0 ,105269*, b_2= 0,1702, b_3= *-0 ,028*)

18 Bejblík
a. b_1= 0,05005198, b_2= 0,01087146, b_3= 0,0737228
b. b_1= 0,03781, b_2= 0,01017, b_3= 0,06929
c. (b_1= 0,1179202, b_2= 0,03582004, b_3= *-0 ,1947484*)
d. (b_1= 0,097001, b_2= 0,038387, b_3= *-0 ,1904*)

19 Jacko
a. b_1= 0, 06296325 , b_2= 0, 05349011 , b_3= 0, 07177231
b. b_1= 0, 04786 , b_2= 0, 07552 , b_3= 0, 07078
c. (b_1= *-0 ,02556444* , b_2= 0,383266, b_3= 0,02590992)
d. (b_1= *-0 ,03408* , b_2= 0,41475, b_3= 0,01842)

20 Petlan
a. (b_1= 0,02114794, b_2= 0,1075671, b_3= *-0 ,002279997*)
b. (b_1= 0,01093, b_2= 0,1306, b_3= *-0 ,003033*)
c. b_1= 0,1003655, b_2= 0,2763939, b_3= 0,03365033
d. b_1= 0,09735, b_2= 0,3285, b_3= 0,03435

Ad přístup II.:

1 Poe
   a. (b_1= *-0 ,002452624*, b_2= 0,08604768, b_3= 0,0685204)
   b. (b_1= *-0 ,01255*, b_2= 0,1192, b_3= 0,06797)
   c. (b_1= *-0 ,09433318*, b_2= 0,2607647, b_3= 0,08967)
   d. (b_1= *-0,118624*, b_2= 0,3274, b_3= 0,08967)

2 Babler – německý
   a. (b_1= *-0 ,009175805*, b_2= 0,01780423, b_3= 0,1174008)
   b. (b_1= *-0 ,02042*, b_2= 0,03187, b_3= 0,1227)
   c. b_1= 0, 008098222, b_2= 0, 1528923, b_3= 0, 2874386
   d. (b_1= *-0 ,01935*, b_2= 0,18678, b_3= 0,2975)

8 Baber – český – použití této metody pro český překlad není smysluplné

Ad přístup III.:

1 Poe
   a. (b_1= 0,0322245, *b_2= -0,01034285*, b_3= 0,08616824)
   b. b_1= 0,02651, b_2= 0,014, b_3= 0,08465
   c. (b_1= *-0,02803661*, b_2= 0,1738866, b_3= *-0,002206851*)
   d. (b_1= *-0,034281*, b_2= 0,22741, b_3= *-0,01568*)

2 Babler – německý
   a. b_1= 0, 01277375, b_2= 0, 00687688, b_3= 0, 1533014
   b. b_1= 0, 007085, b_2= 0, 01291, b_3= 0, 1585
   c. b_1= 0, 0500845, b_2= 0, 1849599, b_3= 0, 2933871
   d. b_1= 0, 037526, b_2= 0, 2091, b_3= 0, 30062

8 Babler – český
   a. b_1= 0,05880817, b_2= 0,0716555, b_3= 0,109372
   b. b_1= 0,04655, b_2= 0,08912, b_3= 0,1115
   c. b_1= 0,1108025, b_2= 0,3782585, b_3= 0,2381214
   d. b_1= 0,083001, b_2= 0,39362, b_3= 0,23396

žurnalistický text

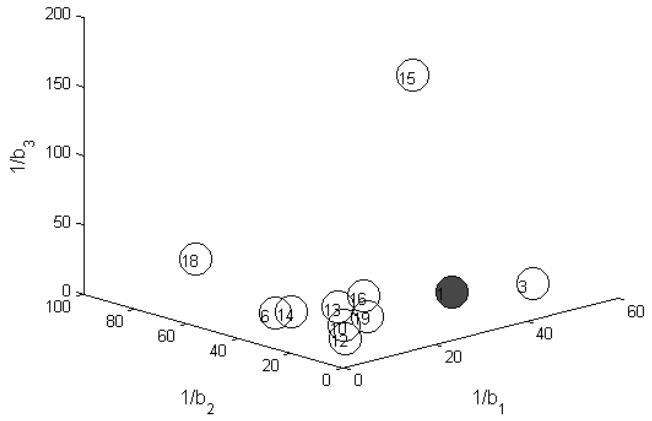
Ad přístup 0.
   a. (b_1= -0,01625, b_2= 0,001512, b_3= *-0,04285*)
   b. (b_1= *-0,02014*, b_2= 0,002468, b_3= *-0,03903*)
   c. (b_1= 0,09209, b_2= 0,30998, b_3= *-0,35815*)
   d. (b_1= 0,08753, b_2= 0,30393, b_3= *-0,3561*)

Ad přístup III.
   a. (b_1= *-0,01014*, b_2= -0,06567, b_3= 0,05374)
   b. (b_1= -0,01303, b_2= *-0,06866*, b_3= 0,05363)
   c. b_1= 0,07906, b_2= 0,18043, b_3= 0,076224
   d. b_1= 0,07311, b_2= 0,17141, b_3= 0,072443

V následujících tabulkách jsou prezentovány reciproké hodnoty parametrů b_1, b_2, b_3 získané z těch výběrových souborů, které se ukázaly být jazykovými fraktály dle definice (tzn., všechny parametry b_1, b_2, b_3 jsou kladné). Následující tabulky jsou obohaceny o reciproké hodnoty aritmetických průměrů $D = \frac{3}{b_1+b_2+b_3}$, kterou nazýváme mírou sémantičnosti příslušných textových výběrových souborů, a dále o pořadí v žebříčku dimenzionality. Pro další detaily viz (Andres, 2009) a (Andres et al., 2011). Výstupy výpočtů jsou seřazeny do pěti tabulek v závislosti na tom, jaký přístup pro stanovování jednotek a jaká metoda pro výpočet parametrů byla použita. Tabulky jsou doprovázeny 3D grafy, které ilustrují pozici bodů se souřadnicemi $\frac{1}{b_1}, \frac{1}{b_2}, \frac{1}{b_3}$, které reprezentují způsob, jakým byl každý jednotlivý výběrový soubor analyzován. Každá tabulka obsahuje sloupec *D* – pořadí, ve kterém jsou výběrovým souborům přiřazena pořadí dle velikosti jejich dimenzí, viz tab. č. 8_a, 8_b, 8_c, 8_d a 9 a obr. č. 9_a, 9_b, 9_c, 9_d a 10 demonstrují vzájemné pozice výstupů kvantifikace.

		$1/b_1$	$1/b_2$	$1/b_3$	*D*	*D* – pořadí
1a	Poe	35,3452	20,3577	14,5942	20,5572	3
3a	Šembera	53,9049	21,4526	4,1072	9,7200	10
6a	Lutinov	10,7899	44,9866	6,6466	11,3054	9
10a	Stoklas	9,8626	16,9097	13,6426	12,8298	8
12a	Havel	10,1535	16,9337	3,8310	7,1673	11
13a	Čapek	16,0353	30,2497	14,1268	18,0496	4
14a	Resler	10,2940	38,0969	11,5119	14,2681	6
15a	Černý	21,6515	11,3035	187,0486	21,4285	2
16a	Slavík	10,7446	11,0534	36,7958	14,2371	7
18a	Bejblík	19,9792	91,9840	13,5643	22,2806	1
19a	Jacko	15,8823	18,6950	13,9329	15,9383	5

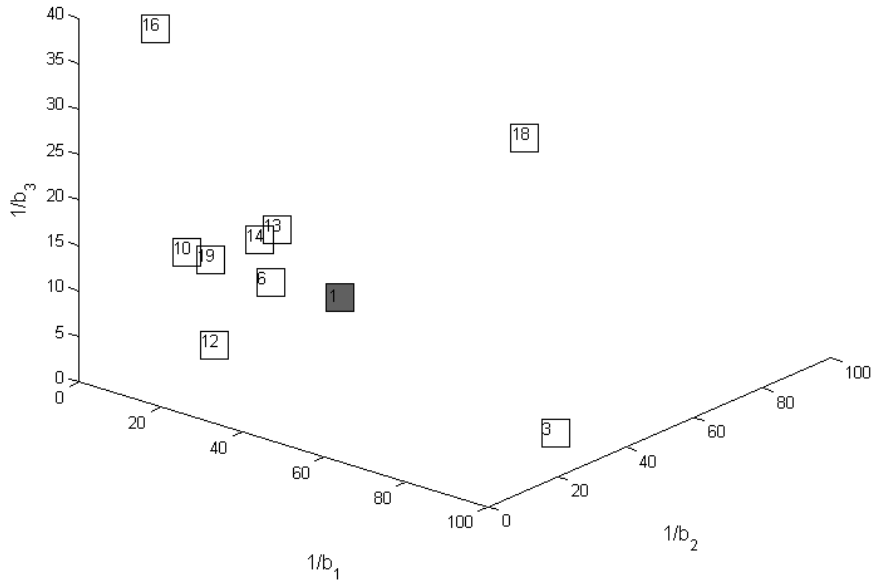
Tab. č. 8_a (přístup I.): Reciproké hodnoty parametrů b_i, i = 1, 2, 3 a jejich aritmetické průměry pro jednoduchou verzi formule MAL získané pomocí metody a I



Obr. č. 9_a: Pozice výstupů kvantifikace prezentovaných v tab. č. 8_a ve 3D (černý kruh odkazuje na anglický originál)

		$1/b_1$	$1/b_2$	$1/b_3$	D	D – pořadí.
1b	Poe	53,7346	11,8301	14,7124	17,5326	3
3b	Šembera	98,0392	22,1190	4,1477	10,1177	9
6b	Lutinov	15,1012	37,8358	6,8399	12,5602	7
10b	Stoklas	13,9606	14,7362	13,8274	14,1636	6
12b	Havel	19,0767	16,5344	4,0388	8,3218	10
13b	Čapek	21,3721	32,1130	14,4928	20,4179	2
14b	Resler	14,7536	35,0140	11,9048	16,6352	4
16b	Slavík	12,3289	7,4349	39,3546	12,4471	8
18b	Bejblík	26,4480	98,3284	14,4321	25,5820	1
19b	Jacko	20,8943	13,2415	14,1283	15,4512	5

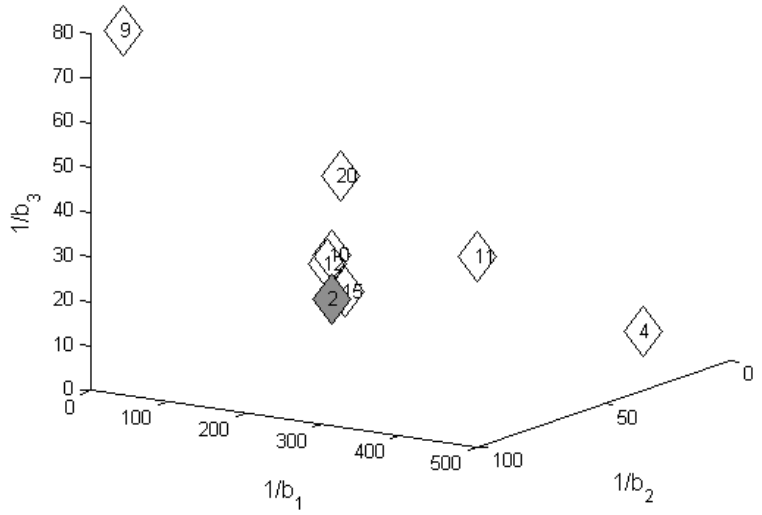
Tab. č. 8_b (přístup I.): Reciproké hodnoty parametrů b_i, i = 1, 2, 3 a jejich aritmetické průměry pro jednoduchou verzi formule MAL získané pomocí metody b I



Obr. č. 9_b: Pozice výstupů kvantifikace prezentovaných v tab. č. 8_b ve 3D (černý čtverec odkazuje na anglický originál)

		$1/b_1$	$1/b_2$	$1/b_3$	D	D – ord.
2c	Babler - německý	7,5667	6,3742	2,6360	4,4884	8
4c	Vrchlický	451,3923	19,3238	9,0593	18,2538	1
9c	Taufer	5,9038	88,4216	78,4883	15,5093	2
10c	Stoklas	5,2699	5,2982	12,2841	6,5230	6
11c	Wagnerová	187,9990	3,8396	16,1885	9,1593	4
12c	Havel	9,2307	8,6269	11,1043	9,5448	3
15c	Černý	20,7793	4,9736	4,2576	6,1976	7
20c	Petlan	9,9636	3,6180	29,7174	7,3098	5

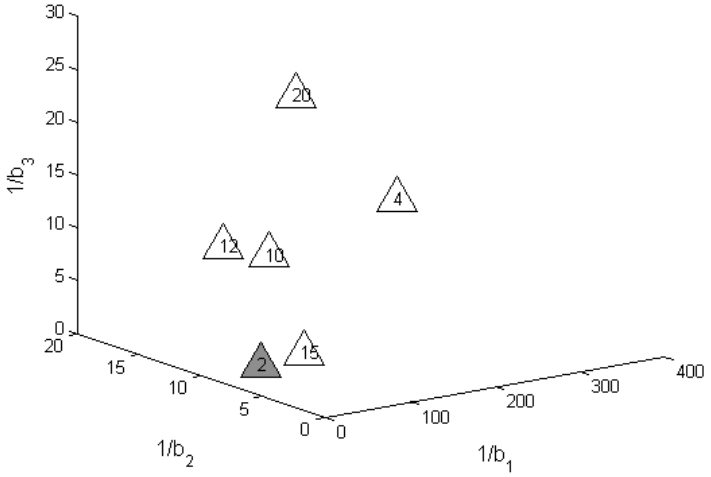
Tab. č. 8_c (přístup I.): Reciproké hodnoty parametrů b_i, i = 1, 2, 3 a jejich aritmetické průměry pro úplnou verzi formule MAL získané pomocí metody c I



Obr. č. 9_c: Pozice výstupů kvantifikace prezentovaných v tab. č. 8_c ve 3D (šedý kosočtverec odkazuje na německý překlad)

		$1/b_1$	$1/b_2$	$1/b_3$	D	D – pořadí.
2d	Babler - německý	8,9884	5,7870	2,5594	4,4459	6
4d	Vrchlický	363,9010	19,1644	7,8846	16,5055	1
10d	Stoklas	5,6196	4,9302	13,3142	6,5806	3
12d	Havel	9,4392	8,8941	12,5063	10,0558	2
15d	Černý	34,7802	4,0967	3,9684	5,7160	5
20d	Petlan	10,2722	3,0441	29,1121	6,5189	4

Tab. č. 8_d (přístup I.): Reciproké hodnoty parametrů b_i, i = 1, 2, 3 a jejich aritmetické průměry pro úplnou verzi formule MAL získané pomocí metody d I

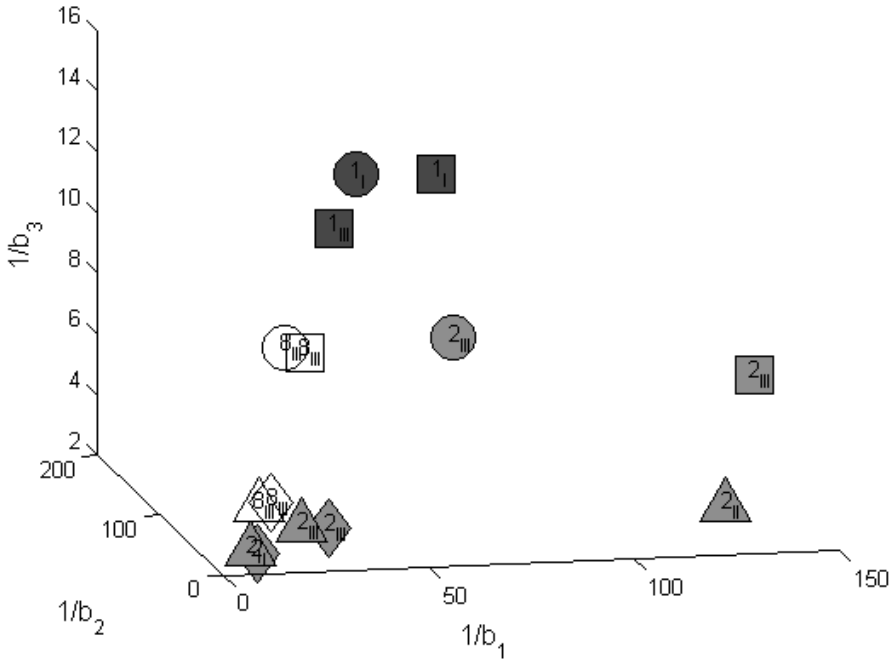


Obr. č. 9_d: Pozice výstupů kvantifikace prezentovaných v tab. č. 8_d ve 3D (šedý trojúhelník odkazuje na německý překlad)

Navzdory tomu, že byl vynechán nevhodný přístup II., se jeví jako smysluplné uvést na obr. č. 10 ještě jeden 3D graf, kde jsou shromážděny body $\left(\frac{1}{b_1}, \frac{1}{b_2}, \frac{1}{b_3}\right)$ získané přístupy I., II., III., které se vztahují k anglickému originálnímu textu (označeno černě), k Bablerovu německému překladu (označeno šedě) a k Bablerovu českému překladu (označeno bíle). Graf demonstruje jejich vzájemný vztah. Primárním důvodem publikování tohoto grafu je exklusivita Otto F. Bablera, který byl překladatelem jak do německého, tak do českého jazyka. Aby byl graf úplný, byl přidán i originální Poeův anglický text, viz tab. č. 9 a obr. č. 10.

		$1/b_1$	$1/b_2$	$1/b_3$	D	D – ord.
1a I	Poe	35,3452	20,3577	14,5942	20,5572	2
1b I	Poe	53,7346	11,8301	14,7124	17,5326	3
1b III	Poe	37,7216	71,4286	11,8133	23,9693	1
2c I	Babler - německý	7,5667	6,3742	2,6360	4,4884	11
2d I	Babler - německý	8,9884	5,7870	2,5594	4,4459	12
2c II	Babler - německý	123,4839	6,5406	3,4790	6,6900	8
2a III	Babler - německý	78,2855	145,4148	6,5231	17,3459	4
2b III	Babler - německý	141,1433	77,4593	6,3091	16,8072	5
2c III	Babler - německý	19,9663	5,4066	3,4085	5,6772	9
2d III	Babler - německý	26,6482	4,7824	3,3265	5,4820	10
8a III	Babler - český	17,0044	13,9557	9,1431	12,5086	6
8b III	Babler - český	21,4823	11,2208	8,9686	12,1374	7
8c III	Babler - český	9,0251	2,6437	4,1995	4,1255	14
8d III	Babler - český	12,0480	2,5405	4,2742	4,2219	13

Tab. č. 9: Reciproké hodnoty parametrů b_i, i = 1, 2, 3 a jejich aritmetické průměry získané pomocí metod a I, b I, c I, d I, c II, a III, b III, c III, d III



Obr. č. 10: Pozice výstupů kvantifikace, které se vztahují k anglickému originálu a Bablerovým překladům

4.2.6 Krok 6 – statistická analýza

Pro demonstraci výše zmíněného postupu si vyberme jako ukázku jeden z textů, který byl již analyzovaný výše – text originálu básně E. A. Poea.

Z výše uvedené regresní analýzy získáme regresní přímku

$$y' = 2{,}461637 + 0{,}02829 \cdot x'.$$

Pro výpočet konfidenčního intervalu použijeme nejprve vzorec pro výpočet reziduálního rozptylu (19)

$$s_r^2 = \frac{S_r}{n-2} = \frac{\sum_{i=1}^{n}(y'_i - a' - b'x'_i)^2}{n-2} = \frac{\sum_{i=1}^{n} y'^2_i - a'\sum_{i=1}^{n} y'_i - b'\sum_{i=1}^{n} x'_i y'_i}{n-2}$$

$$s_r^2 = \frac{109{,}4674 - 2{,}461637 \cdot 45{,}52832 - 0{,}02829 \cdot 104{,}691}{19-2}$$

$$s_r^2 = 0{,}020891$$

$$s_r = 0{,}141536$$

Dále spočteme

$$s_b = s_r \sqrt{\frac{n}{n\sum_{i=1}^{n} x'^2_i - \left(\sum_{i=1}^{n} x'_i\right)^2}}$$

$$s_b = 0.141536 \cdot \sqrt{\frac{19}{19 \cdot 121.5524 - 1929.5}} = 0{,}032319.$$

V tabulkách kritických hodnot Studentova rozdělení najdeme $t_{0{,}05(19-2)} = 2{,}110$. Hledaný konfidenční interval má tvar

$$\left(b - t_{\alpha(n-2)} \cdot s_b, b + t_{\alpha(n-2)} \cdot s_b\right), \tag{51}$$

tedy po dosazení

$$(-0{,}02829 - 2{,}110 \cdot 0{,}032319; -0{,}02829 + 2{,}110 \cdot 0{,}032319)$$

$$(-0{,}0399; 0{,}096486)$$

Je 95% konfidenční interval pro regresní koeficient *β*.

Pro tyto výpočty lze samozřejmě mnohem efektivněji použít počítačový software. Použití budeme opět demonstrovat na software R a jeho již výše zmíněném výstupu, z něhož uvedeme nyní jen relevantní sekci, ve které je patrný konfidenční interval (červeně jsou zvýrazněny parametr *A'*, *b*, modře konfidenční interval pro parametr*A'*, zeleně pro parametr *b*).

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.38656 -0.04728 -0.03152  0.05852  0.32448

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.46164    0.08175  30.113 3.43e-16 ***
lnX         -0.02829    0.03232  -0.875    0.394
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1445 on 17 degrees of freedom
Multiple R-squared: 0.04313,    Adjusted R-squared: -0.01315
F-statistic: 0.7663 on 1 and 17 DF,  p-value: 0.3936

> koef=vysledek$coefficients
> exp(koef[1]);-koef[2]
(Intercept)
  11.72398
      lnX
0.02829237
> # konfidencni intervaly
> koefStd=confint(vysledek)
> exp(koefStd[1,]);-koefStd[2,]
    2.5 %    97.5 %
 9.866719 13.930851
     2.5 %     97.5 %
 0.09648018 -0.03989544
```

V tabulce regresní analýzy ve sloupci *t-value* můžeme číst testová kritéria, která slouží k testování statistické významnosti obou koeficientů regresní přímky a umožňují testování nulové hypotézy. Platnost hypotéz lze posoudit podle posledního sloupce, kde najdeme hladiny významnosti, které by měly být menší než $\alpha = 0{,}05$, aby mohly být odhadnuté

parametry považovány za statisticky významné. V našem případě tedy koeficient *b* statisticky významný není.

Tedy pro tento ukázkový výběrový soubor je konfidenční interval pro parametr b_3 pro jednoduchou verzi formule MAL, kde data jsou zpracována přístupem I., (-0,0399; 0,0965). Je tudíž zjevné, že odhad není dostatečně přesný, neboť konfidenční interval je příliš široký a pokrývá hodnotu nula[32]. Důvod takto nedostatečného odhadu může být např. špatná volba modelu (logaritmická transformace + lineární model). 95%-konfidenční intervaly pro všechny výběrové soubory zpracované přístupy I., II., III. a statistickými metodami a a c jsou uvedeny v příloze VII., zde pro ukázku konfidenční intervaly pro parametry b_i získané pomocí přístupu III., viz tab. č. 10.

Ad přístup III.

		b_1	b_2	b_3
2a	Babler–Ger.	(-0.0644; 0.0899)	(-0.0719; 0.0857)	**(0.0767; 0.2299)**
2c	Babler–Ger.	(-0.1009; 0.2010)	**(0.0559; 0.3141)**	(-0.0434; 0.6302)
8a	Babler–Cz.	(-0.0699; 0.1875)	(-0.0300; 0.1733)	**(0.0044; 0.2144)**
8c	Babler–Cz.	(-0.1724; 0.3940)	**(0.1448; 0.6117)**	(-0.4036; 0.8798)

Tab. č. 10: 95%-konfidenční intervaly parametrů b_1, b_2, b_3 pro výběrové soubory zpracované pomocí přístup III. (intervaly, které nepokrývají hodnotu nula jsou zvýrazněny tučně)

Bohužel u všech výše zmíněných výběrových souborů konfidenční intervaly obsahují hodnotu nula alespoň u jednoho z parametrů b_1, b_2, b_3. Tento fakt může být interpretován tím způsobem, že hodnoty parametrů b_i překračují hodnotu nula a dostávají se do záporných čísel jen velice těsně pod hodnotu nula. V tab. č. 11 jsou tudíž prezentovány hodnoty upravených konfidenčních intervalů s nejnižšími možnými pravděpodobnostní hodnotami tak, aby obsahovaly výhradně kladné hodnoty, dále viz (Andres & Benešová, 2011).

Některé z upravených konfidenčních intervalů stále ještě nejsou vyhovující. Navíc výsledky pro ostatní výběrové soubory jsou podobné nebo ještě horší. Nicméně je nutné si uvědomit, že byla použita metoda linearizace pro nalezení parametrů, které vedly k takovýmto výstupům. Na druhé straně numerické metody jsou aplikovány na nelineární modely s dostatečnou přesností. Tudíž se konfidenční intervaly nestávají tak závažným břemenem naší analýzy.

		b_1		b_2	
Poe	1a I	60%	(0.0004; 0.0562)	70%	(0.0091; 0.0892)
Babler – German	2c I	80%	(0.0164; 0.2479)	95%	**(0.0235; 0.2903)**
Babler – German	2c II	10%	not available	90%	(0.0167; 0.2891)
Babler – German	2a III	20%	(0.0034; 0.0222)	10%	(0.0021; 0.0116)
Babler – German	2c III	50%	(0.0012; 0.0990)	95%	**(0.0559; 0.3141)**

[32] V předpokladech, které musí splňovat jazykový fraktál, bylo stanoveno, že všechny parametry b_i musí být nutně kladné.

Babler – Czech	8a III	60%	(0.0065; 0.1111)	80%	(0.0084; 0.1349)
Babler – German	8c III	50%	(0.0194; 0.2022)	95%	**(0.1448; 0.6117)**

			b_3
Poe	1a I	70%	(0.0015; 0.1355)
Babler – German	2c I	95%	**(0.2028; 0.5559)**
Babler – German	2c II	80%	(0.3381; 0.5411)
Babler – German	2a III	95%	**(0.0767; 0.2299)**
Babler – German	2c III	90%	(0.0648; 0.5219)
Babler – Czech	8a III	95%	**(0.0044; 0.2144)**
Babler – German	8c III	70%	(0.0314; 0.4448)

Tab. č. 11: Upravené konfidenční intervaly s výhradně kladnými hodnotami parametrů b_1, b_2, b_3 (jsou doplněné svou příslušnou nejvyšší možnou pravděpodobnostní hodnotou). Původní 95%- konfidenční intervaly jsou zvýrazněné tučně.

Koeficient determinace

Pro výpočet „těsnosti“ modelu získaného pomocí statistických metod, použijeme koeficient determinace. Tedy pokud uvažujeme data v tab. č. 1_3 a logaritmizujeme, dostáváme:

	x_3	z_3	y_3	ln x_3	ln y_3	ln $y_{3\,est}$	$\sum_{i=1}^{n}(y_j - y_{j\,est})^2$	$\sum_{i=1}^{n}(y_j - \bar{y})^2$
	1	115	2,4870	0	0,911077	0,915157	1,66442E-05	0,003051
	2	181	2,4392	0,693147	0,89167	0,879555	0,000146778	0,001284
	3	176	2,3542	1,098612	0,856201	0,858729	6,39138E-06	1,31E-07
	4	108	2,2963	1,386294	0,831299	0,843953	0,000160119	0,000602
	5	30	2,2200	1,609438	0,797507	0,832492	0,001223912	0,003403
	6	2	2,3333	1,791759	0,847284	0,823127	0,000583535	7,32E-05
∑					5,135038		0,002137379	0,008413

Tedy

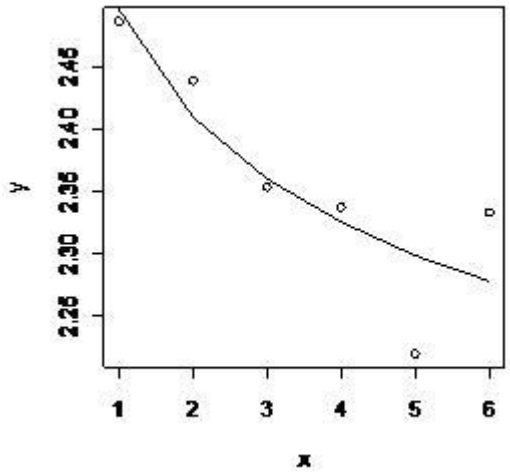
$$R^2 \doteq 1 - \frac{0{,}002137379}{0{,}008413}$$
$$R^2 \doteq 0{,}745951.$$

Pro snadnější a efektivnější výpočet koeficientu determinace je samozřejmě opět výhodnější použít statistický software, v případě tohoto experimentu volím opět R software. Tato hodnota může být z modelu získána funkcí `summary()` hodnoty `r.squared`.

```
> summary(model1)$r.squared

[1] 0.7426771
```

```
> summary(model2)$r.squared

[1] 0.7444607[33]
```

Koeficient determinace je tedy roven přibližně 0,7444607 v modelu pro jednoduchou verzi formule MAL. Interval, ve kterém se koeficient determinace může pohybovat, je $0 \leq R^2 \leq 1$. Čím blíže je koeficient k 1, tím lépe model sedí, viz obr. č. 11. Hodnoty R^2 větší nebo rovny 0,7 mohou prokazovat adekvátní a dobře sedící model v kvantitativní lingvistice. Hodnota R^2 = 0,7 může být interpretována jako fakt, že regresním modelem je vysvětlena 70% variabilita hodnot *y*, viz (Heibeger & Holland, 2004).



Obr. č. 11: Regresní křivka a izolované body znázorňující empirické hodnoty z tab.č. 43_3 v příloze I.

4.2.7 Krok 7 – fraktální analýza

V případě našeho naposledy zmíněného výběrového souboru vezmeme v úvahu hodnoty parametrů $A_i, b_i, c_i, i = 1,2,3$, z tab. č. 43_3 a vezmeme $k = 14$ jako nejmenší kladné celé číslo větší než

$$\max\left(\frac{1}{b_1}, \frac{1}{b_2}, \frac{1}{b_3}\right) = \max(13{,}67802\ldots;\ 5{,}833965\ldots;\ 13{,}80396\ldots) = 13{,}80396\ldots.$$

Pro číslo $m = x^k$ kontrakcí ${}_if_j$ v (53) takto dostaneme $m = x^{14}$, tj. $m = 2^{14} = 16\,384$, pro $x = 2$, a $m = 3^{14} = 4\,782\,969$, pro $x = 3$, atd.

Například, pro $x = 2$, můžeme také snad vypočítat faktory kontrakce r_1, r_2, r_3 v (53) jako $r_1 = \frac{1}{2^{14\,.0{,}07311}} = 0{,}4919\ldots, r_2 = \frac{1}{2^{14\,.0{,}17141}} = 0{,}1895\ldots, r_3 = \frac{1}{2^{14\,.0{,}072443}} = 0{,}4951\ldots$

Fraktální dimenze *D* fraktálu $\mathbf{A} = F(\mathbf{A})$, kde *F* je definováno v (53), může být spočteno, s ohledem na (26), jako $D = \frac{3}{b_1+b_2+b_3} = 9{,}464827\ldots$, a pro 2D a 3D projekce máme $D^{(2)} = \frac{D}{2} = 1{,}352118\ldots$, $D^{(3)} = \frac{3}{14}D = 2{,}028177\ldots$.

Samotný fraktál $\mathbf{A}$ může být generován pomocí (53) a (54) bere podobu

[33] Odchylka od výsledku mechanického výpočtu je způsobena zaokrouhlováním.

$$d_H(F^s([0,1]),\mathbf{A}) \leq \frac{\left(\frac{1}{2}+\frac{\sqrt{13}}{2}.0.1895\ldots\ 0.4951\ldots\right)}{2^{S14\,.\ 0.316963\ldots}}/(1-2^{-14\,.\ 0.316963\ldots}) \leq \frac{0.4919\ldots0.1895\ldots0.4951\ldots}{1-0.4919\ldots0.1895\ldots0.4951\ldots}\sqrt{14}.$$

Speciálně dostaneme $d_H(F([0,1]),\mathbf{A}) \leq 0{,}0609\ldots$ Protože je to už dostatečně malé číslo pro optické rozlišení, $F([0,1])$ může být považován za model zkoumané textové struktury. Poznamenejme, že méně přesný odhad v (54) dává výsledek pouze $d_H(F([0,1]),\boldsymbol{A}) \leq 0,,181033\ldots$, což by bylo nevhodné pro naše potřeby, tedy abychom považovali $F([0,1])$ za model.

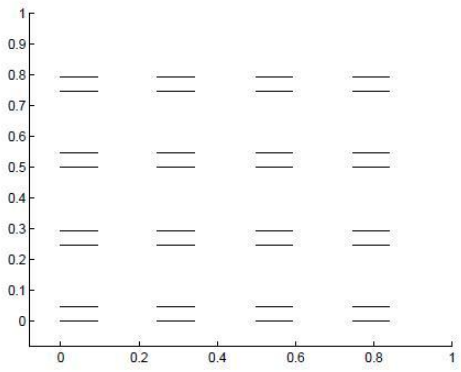
4.2.8 Krok 8 – vizualizace

Jako první příklad pro vizualizaci si zvolme opět data z tab. č. 43_3. Faktory kontrakce pro $x = 2$ jsou následující

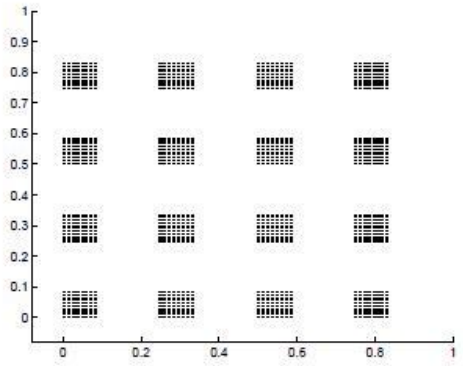
$$r_1 = \frac{1}{2^{14\,.0{,}7311}} \doteq 0{,}4919$$

$$r_2 = \frac{1}{2^{14\,.0{,}17141}} \doteq 0{,}1895$$

$$r_3 = \frac{1}{2^{14\,.0{,}072444}} \doteq 0{,}4951.$$



Obr. č. 12_1: Dvojdimenzionální projekce první aproximace ***A***
(vizualizace modelu jazykového fraktálu pro žurnalistický text)



Obr. č. 12_2: Dvojdimenzionální projekce druhé aproximace ***A.***
(vizualizace přidruženého matematického fraktálu k fraktálu jazykovému z obr. č. 12_1)

Připomeňme si, jazykovými fraktály máme na mysli právě ty výběrové soubory, jejichž parametry b_1, b_2, b_3 jsou kladné a splňují MAL, viz (Andres, 2009), (Andres, 2010) a (Andres et al., 2011).

Fraktální analýza byla částečně provedena v předcházejících krocích, kde jsme přiřadili každému výběrovému souboru bod $\left(\frac{1}{b_1},\frac{1}{b_2},\frac{1}{b_3}\right)$ ve trojrozměrném Eukleidovském prostoru a speciálně též hodnota $D = \frac{3}{b_1+b_2+b_3}$.

Tudíž se v několika následujících odstavcích zaměřím na vizualizaci některých modelů dalších význačných výběrových souborů. V podstatě se omezím na jazykové fraktály 3. řádu, viz níže. Z tohoto důvodu použijeme univerzální konstrukci popsanou výše a v (Andres & Rypka, 2011) a (Andres et al., 2011).

Každá vizualizace výběrového souboru je, jak bylo řečeno výše, její dvojdimenzionální projekcí z prostoru, jehož celočíselná dimenze je větší nebo rovna maximu z $\frac{1}{b_1},\frac{1}{b_2},\frac{1}{b_3}$.

Kvůli možným komparacím by bylo optimální provádět projekce z prostoru o stejné dimenzi pro všechny výběrové soubory, tj. v našem případě z prostoru o dimenzi 452[34]. Bohužel by ale při takovéto projekci nebylo u mnoha výběrových souborů možné rozlišit detaily vizualizace, protože tyto by byly redukovány na pouhých několik bodů (tak je tomu např. v případě výběrových souborů 1b III, 2c II, 2a III, 2b III, 3b I, 4c I, 4d I, 9c I, 11c I, 15a I, 18a I, 18b I). Takové projekce nemá valný význam vizualizovat. Možná ne tak drastická, ale v principu podobná situace nastává v případě dvojdimenzionálních projekcí z Eukleidovských prostorů s dimenzemi 188 (metoda a I.), 99 (metoda b I.), 452 (metoda c I.), 364 (metoda d I.), 124 (metoda c II.), 79 (metoda a III.), 142 (metoda b III.), 20 (metoda c III.) a 27 (metoda d III.), pokud aplikujeme metody odděleně.

Na druhou stranu by k těmto problémům nedošlo, pokud bychom projekce provedli z prostoru s dimenzí, která je maximálně o 1 vyšší, než je celá část z $\frac{1}{b_1},\frac{1}{b_2},\frac{1}{b_3}$. V případě

[34] Maximální převrácená hodnota všech parametrů b_i ze všech výběrových souborů je b_1 = 451,3923 u Vrchlického českého překladu 4c I.

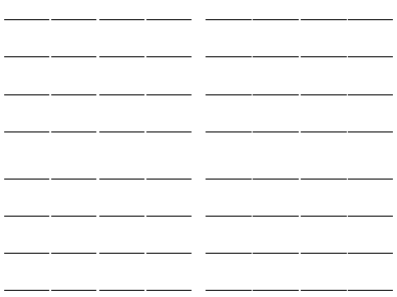
takovéto analýzy nejsou vizualizace obvykle redukovány na pouhé body, jak je patrné např. z obr. č. 13_1. Bohužel je ale negativem takové analýzy to, že nemůžeme poté srovnávat vizualizace jednotlivých výběrových souborů. V textu jsou prezentovány vizualizace jazykových fraktálů 3. řádu. Ostatní vizualizace viz příloha VIII.

Obrázky č. 13_2, 14_2 a 15_2 uvedené zde se týkají přidružených matematických fraktálů. Aproximace jejich modelů jsou jazykové fraktály 13_1, 14_1 a 15_1.

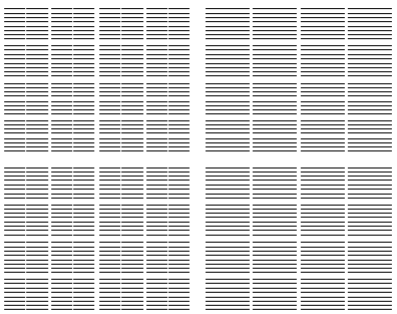
Již zmíněné speciální typy jazykových fraktálů jsou ty, jejichž dva nebo více parametrů b_1, b_2, b_3 jsou si přibližně rovny. Nazývejme takové objekty jazykovými fraktály 2. a 3. řádu, v tomto pořadí.

Jazykové fraktály 3. řádu

Stoklas (obr. č. 13)	10b I	$\lvert b_{max} - b_{min}\rvert = b_3 - b_2 = 0{,}00446$	$b \doteq 0{,}07$
Jacko (obr. č. 14)	19a I	$\lvert b_{max} - b_{min}\rvert = b_3 - b_2 = 0{,}0182822$	$b \doteq 0{,}06$
Havel (obr. č. 15)	12c I	$\lvert b_{max} - b_{min}\rvert = b_2 - b_3 = 0{,}0258611$	$b \doteq 0{,}10$

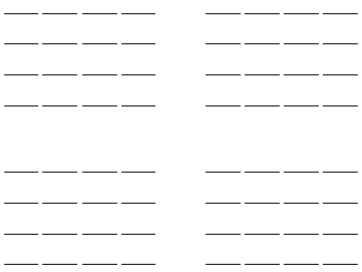


Obr. č. 13_1: se Vizualizace modelu jazykového fraktálu vztahujícímu ke Stoklasovu překladu 10b I. (dvojdimenzionální projekce z prostoru s dimenzí 15)

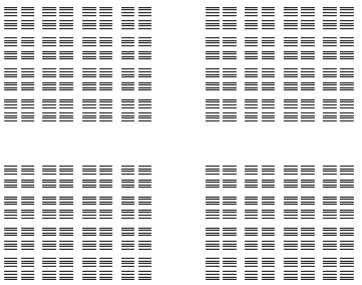


Obr. č. 13_2: Vizualizace přidruženého matematického fraktálu, jehož aproximací je jazykový fraktál na obr. č. 13_1 (dimenze jeho dvojdimenzionální projekce je $D^{(2)} \doteq 1{,}88848493$[35])

[35] Dimenzi dvojdimenzionální projekce matematického fraktálu přidruženého k fraktálu jazykovému lze spočítat následujícím způsobem:



Obr. č. 14_1: se Vizualizace modelu jazykového fraktálu vztahujícímu k Jackovu překladu 19a I. (dvojdimenzionální projekce z prostoru s dimenzí 19)



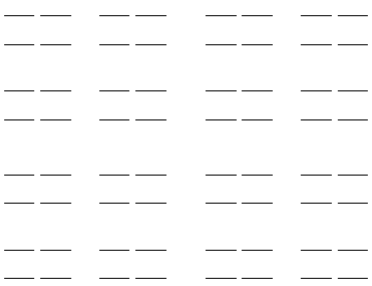
Obr. č. 14_2: Vizualizace přidruženého matematického fraktálu, jehož aproximací je jazykový fraktál na obr. č. 14_1 (dimenze jeho dvojdimenzionální projekce je $D^{(2)} \doteq 1.67771737$)

$$D^{(2)} = \frac{2}{k} \cdot D,$$

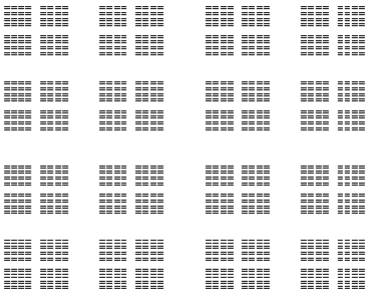
kde *k* je dimenze, se kterou začínáme v případě fraktálu jazykového, *D* je míra sémantičnosti jazykového fraktálu, jak bylo definováno výše a např. v (Andres, 2009).

V konkrétním případě matematického fraktálu na obr. č. 13_2 přidruženého k jazykovému fraktálu na obr. č. 13_1 dostaneme:

$$D^{(2)} = \frac{2}{15} \cdot 14{,}163637 \doteq 1{,}8884849.$$



Obr. č. 15_1: se Vizualizace modelu jazykového fraktálu vztahujícímu k Havlovu překladu 12c I. (dvojdimenzionální projekce z prostoru s dimenzí 12)



Obr. č. 15_2: Vizualizace přidruženého matematického fraktálu, jehož aproximací je jazykový fraktál na obr. č. 15_1 (dimenze jeho dvojdimenzionální projekce je $D^{(2)} \doteq 1.5908057$)

Jazykové fraktály 2. řádu:

Stoklas	10c I	$b_1 - b_2 = 0{,}00101$	$b = b_1 \doteq b_2 \doteq 0{,}18$
Taufer	9c I	$b_3 - b_2 = 0{,}0014$	$b = b_2 \doteq b_3 \doteq 0{,}01$
Slavík	16a I	$b_1 - b_2 = 0{,}0026$	$b = b_1 \doteq b_2 \doteq 0{,}09$
Jacko	19b I	$b_2 - b_3 = 0{,}00474$	$b = b_2 \doteq b_3 \doteq 0{,}07$
Havel	12d I	$b_2 - b_3 = 0{,}0065$	$b = b_2 \doteq b_3 \doteq 0{,}11$
Černý	15d I	$b_3 - b_2 = 0{,}0079$	$b = b_2 \doteq b_3 \doteq 0{,}248$
Čapek	13a I	$b_3 - b_1 = 0{,}0084$	$b = b_1 \doteq b_3 \doteq 0{,}0656$
Resler	14a I	$b_1 - b_3 = 0{,}0103$	$b = b_1 \doteq b_3 \doteq 0{,}092$
Poe	1b III	$b_1 - b_2 = 0{,}013$	$b = b_1 \doteq b_2 \doteq 0{,}02$
Čapek	13b I	$b_1 - b_2 = 0{,}016$	$b = b_1 \doteq b_1 \doteq 0{,}039$
Babler-německý	2c I	$b_2 - b_1 = 0{,}025$	$b = b_1 \doteq b_2 \doteq 0{,}14$
Černý	15c I	$b_3 - b_2 = 0{,}034$	$b = b_2 \doteq b_3 \doteq 0{,}218$

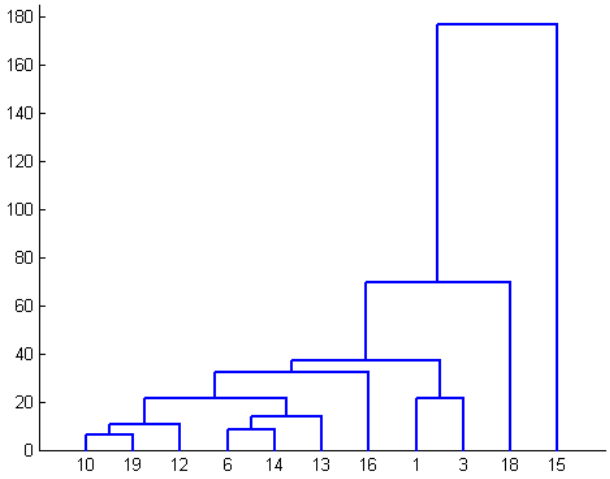
Vizualizace ostatních modelů jazykových fraktálů a k nim přidružených matematických fraktálů jsou dotupné v příloze VIII. Pro tuto přílohu byly vybrány modely, jež nebyly

redukovány na několik nezřetelných bodů a jsou při daném rozlišení oka a monitoru alespoň minimálně patrné.

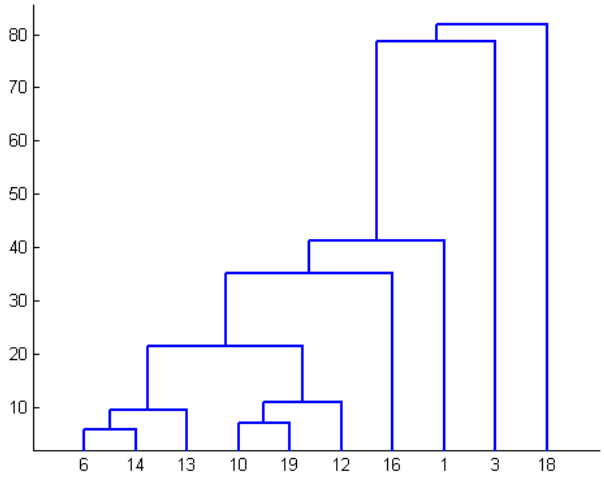
4.2.8.1 Shluková analýza

Na obr. č. 16 a 17 jsou znázorněny dendrogramy vztahující se k obr. č. 9 a 10 (dokonce i tyto původní 3D grafy heuristicky signalizují, že se celá původní množina objektů rozpadá na jednotlivé shluky). Na horizontálních osách jsou vyznačeny notační symboly jednotlivých výběrových souborů, zatímco na osách vertikálních jsou indikovány Eukleidovské vzdálenosti mezi nejbližšími shluky.

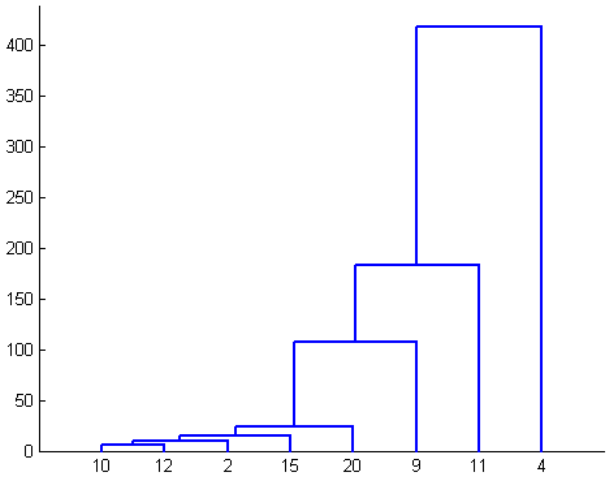
Použití dendrogramů se zdá být naprosto optimální, abychom tak mohli demonstrovat vzájemnou blízkost a souvztažnost v rámci naší analýzy mezi jednotlivými výběrovými soubory. Dendrogramy neukazují pouze Eukleidovské vzdálenosti mezi shluky výběrových souborů, ale také citlivost aplikovaných technik. Abychom byli konkrétnější, z obr. č. 16_a a 16_b je patrné, že existují dvě trojice výběrových souborů (označených jako 10 – 19 – 12 a 6 – 14 – 13) a jeden pár výběrových souborů (označených jako 1 – 3), jejichž objekty jsou si nejbližší. Všechny jejich Eukleidovské vzdálenosti jsou menší než 20, atd. Na druhé straně výběrový soubor, který má od ostatních výběrových souborů největší Eukleidovskou vzdálenost, je označen číslem 15. Další detaily viz (Andres & Benešová, 2011).



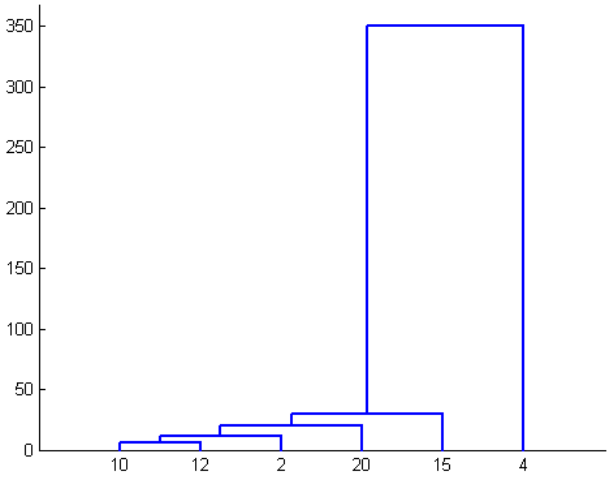
Obr. č. 16_a: Dendrogram vztahující se k obr. č. 9_a



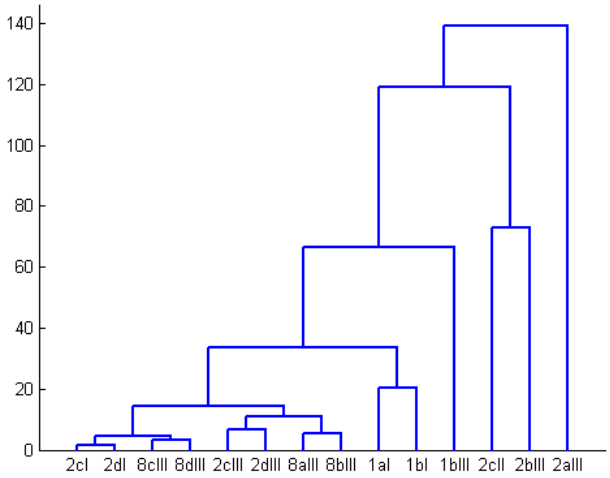
Obr. č. 16_b: Dendrogram vztahující se k obr. č. 9_b



Obr. č. 16_c: Dendrogram vztahující se k obr. č. 9_c



Obr. č. 16_d: Dendrogram vztahující se k obr. č. 9_d



Obr. č. 17: Dendrogram vztahující se k obr. č. 10

Jak již bylo zmíněno jednou výše, obr. 16_a a 16_b znovu dokazují, že metody a a b korespondují. Oba dva grafy znázorňují, že shluky na nejnižší úrovni jsou tvořené jeden Stokalsovým a Jackovým překladem a druhý Lutinovovým a Reslerovým překladem, tzn., tyto překlady jsou si nejbližší. Mezi shluky na druhé úrovni jsou zahrnuty jednou Stoklasův, Jackův a Havlův překlad a po druhé Lutinovův, Reslerův a Čapkův překlad. V této souvislosti je velice pozoruhodné, že Stoklasův, Havlův a Jackův překlad vykazují vlastnosti fraktálů 3. řádu. Překlady, které jsou od ostatních nejvzdálenější, jsou Bejblíkův a Černého.

Obr. č. 16_c a 16_d ilustrují metody c a d, které opět korespondují. Shluky na nejnižší úrovni jsou tvořeny Stoklasovým a Havlovým překladem. Na nejbližší vyšší úrovni k nim přibývá Bablerův překlad do německého jazyka, který je následovaný Černého a Petlanovým překladem na dalších dvou hladinách v tomto pořadí v případě metody c a v pořadí opačném v případě metody d.

Dendrogram na obr. č. 17 odráží skutečnost, že mezi nejbližšími si výběrovými soubory jsou překlady od jednoho autora (v našem případě jde o překlady Otto. F. Bablera do českého a německého jazyka) bez ohledu na metodu nebo přístup, který je zvolen.

4.2.9 Krok 9 – interpretace získaných výsledků analýzy

Tento experiment je primárně zaměřen na následující problémy: zaprvé je nutné řešit segmentaci výběrových souborů, tj. volit jednotky efektivně a zároveň lingvisticky korektně. Nástroje kvantitativní analýzy, které byly zvoleny pro tento experiment, jsou zcela nové a experimentálně ne zcela dostatečně prověřené, proto jsem na úvod zvolila přístup 0., který je nejméně namáhavý a časově náročný, ale na druhé straně ne zcela efektivní a lingvisticky korektní. Jedná se o přístup, kterým byl celý experiment zahájen a s jehož pomocí byla budována metodologie, která nutně musela být nejprve nastavena. Výstup tohoto přístupu nebyl v porovnání s přístupem I. a obzvlášť s přístupy II. a III. uspokojivý. Všechny čtyři přístupy na originální Poeův anglický text básně *The Raven*, přístup II. navíc na Bablerův německý překlad, přístup III. navíc na Bablerův český překlad a přístup I. na všechny momentálně zvolené výběrové soubory. Přístup II. se experimentálně ukázal být naprosto neefektivní i pro

typy jazyků, pro které může vůbec být použit. Tento fakt podpořil teorii, že členy jsou považovány za samostatné jednotky při segmentování výběrových souborů. Jako nejoptimálnější se ukázal přístup III. zejména u obou Bablerových překladů. Oba dva tyto výběrové soubory poskytly požadované výsledky při použití všech čtyř výše zmíněných metod pro nalezení potřebných parametrů, tedy oba výběrové soubory se bez ohledu na použitou metodu ukázaly být jazykovými fraktály.

Babler – německý 2 III	a	$D \doteq 17{,}346$
	b	$D \doteq 16{,}807$
	c	$D \doteq 5{,}677$
	d	$D \doteq 5{,}482$
Babler – český 8 III	a	$D \doteq 12{,}509$
	b	$D \doteq 12{,}137$
	c	$D \doteq 4{,}126$
	d	$D \doteq 4{,}222$
(Poe 1 III	b	$D \doteq 23{,}969$)

Výsledky pro metody a, b, c, d pro oba překlady korespondují. Oba dva sety výsledků zároveň ukazují, že metoda pro výpočet výsledků a koresponduje s metodou b a c koresponduje s d. Zároveň nám metoda c poskytuje nejlepší výsledky. Tato zmíněná fakta jsou prokázána také shlukovou analýzou, která byla demonstrována v dendrogramech. Metody a, b jsou komplementární k metodám c, d, na druhou stranu ale významně mění výsledky analýzy co do velikosti. Z lingvistického hlediska se ukázal jako nejefektivnější přístup III., z formálního hlediska potom metoda c vyšla z porovnání jako nejlepší.

Na druhou stranu výsledky experimentu pro originální anglický Poeův text se nechovají stejným způsobem. Není zřejmě možné a hlavně vhodné hledat zdůvodnění v exkluzivitě a originalitě tohoto výběrového souboru, je třeba zůstat oběma nohama na zemi. Všechny záporné hodnoty parametrů, které odporují definici jazykového fraktálu, jsou velice blízké nule, takže důvodem může být potenciální chyba.

Poe 1 III	a	$b_2 \doteq -0{,}01034285$
	c	$b_1 \doteq -0{,}02803661$, $b_3 \doteq -0{,}002206851$
	d	$b_1 \doteq -0{,}034281$, $b_3 \doteq -0{,}01568$

Následující přehled přináší hodnoty nejbližší a na druhé straně nejvzdálenější k hodnotám získaným kvantifikací originálního anglického Poeova textu. Hodnoty jsou roztříděny dle použitého přístupu a metody.

Hodnoty nejbližší originálnímu Poeovu textu:

Poe 1 I	a	Černý 15a I	$\Delta \doteq 0{,}871$
		Bejblík 18a I	$\Delta \doteq 1{,}723$
		Čapek 13a I	$\Delta \doteq 2{,}507$

	b	Babler – německý 2b III	$\Delta \doteq 0{,}726$
		Resler 14b I	$\Delta \doteq 0{,}898$
		Čapek 13b I	$\Delta \doteq 2{,}885$
Poe 1 III	b	Bejblík 18b I	$\Delta \doteq 1{,}613$
		Čapek 13 b I	$\Delta \doteq 3{,}551$
		Poe 1b I	$\Delta \doteq 6{,}436$

Hodnoty nejvzdálenější od originálního Poeova textu:

Poe 1 I	a	Havel 12a I	$\Delta \doteq 13{,}390$
		Šembera 3a I	$\Delta \doteq 10{,}837$
		Lutinov 6a I	$\Delta \doteq 9{,}252$
	b	Havel 12b I	$\Delta \doteq 9{,}211$
		Bejblík 18b I	$\Delta \doteq 8{,}049$
		Šembera 3b I	$\Delta \doteq 7{,}415$
Poe 1 III	b	Havel 12b I	$\Delta \doteq 15{,}647$
		Šembera 3b I	$\Delta \doteq 13{,}851$
		Slavík 16b I	$\Delta \doteq 11{,}086$

V následujícím přehledu jsou porovnány jednotlivé metody a především je demonstrováno, kolik parametrů b_1, b_2, b_3 bylo záporných, tj. které binarismy by v dalších experimentech mohly být revidovány.

	metoda a	metoda b	metoda c	metoda d	celkově záporných b_i
b_1	6	6	5	7	24
b_2	4	4	0	0	8
b_3	1	2	9	10	22

Právě zmíněná statistika ukazuje, jak těsně jsou spjaty metody a s b a c s d.

Za druhé jsem zamýšlela testovat Menzerath-Altmannův zákon na různých výběrových souborech ve třech různých jazycích, které ale mají stejné sémantické pozadí. Možnost porovnat výsledky získané kvantifikací výběrových souborů od jednoho autora v různých jazycích se ukázala být velkou příležitostí. Porovnání Bablerových překladů do českého a německého jazyka bylo již výše zmíněno. Jeden překlad velice pozoruhodně odráží druhý překlad.

Stupeň sémantičnosti pro Bablerův překlad do němčiny je v případě všech čtyř metod a III., b III., c III. a d III. větší než stupeň sémantičnosti pro Bablerův překlad do češtiny. V případě metody b III. u anglického originálu Poeova *The Raven* je stupeň sémantičnosti vyšší než v případě obou Bablerových překladů. Tudíž se zdá, že stupeň sémantičnosti pro výběrový soubor originálního anglického Poeova *The Raven* je větší než pro jeho německé mutace, jejichž stupeň sémantičnosti je zároveň větší než pro jeho české mutace. Takovýto závěr ale

není udržitelný s ohledem na fakt, že hodnoty *D* v jednotlivých tabulkách vykazují relativně velký rozptyl $D_{\max} - D_{\min}$:

15.113353	in Table 1_a (a I)
17.260215	in Table 1_b (b I)
13.765495	in Table 1_c (c I)
12.05953	in Table 1_d (d I)
19.843807	in Table 2.

Za třetí jsem zamýšlela otestovat textové výběrové soubory na fraktalitu. Aby se prokázala fraktalita textového výběrového souboru, musí být splněny dva výše zmíněné podmínky. Byly vizualizovány nejen jazykové fraktály vyššího řádu, k vizualizacím jazykových fraktálů byly připojeny i vizualizace k nim přidružených matematických fraktálů. Afinita některých výběrových souborů byla vizualizovaná pomocí dendrogramů, které byly sestaveny pomocí shlukové analýzy.

Všechny mezivýpočty, parametry *A*, *c*, grafy, tabulky a výstupy statistického software, které nebyly prezentovány v textu, jsou dostupné v přílohách.

Nicméně je zcela nutné poznamenat, že tato práce, je společně s (Andres, 2009), (Andres, 2010), (Andres et al., 2011) a (Andres & Benešová, 2011) první z analýz tohoto odvětví kvantitativní analýzy. Proto neaspiruje na to, aby prezentovala jakékoli lingvistické univerzálie. Alespoň ne v této fáze experimentů. Celý výzkum si žádá celé množství dalších experimentů, aby byly dokázány výše zmíněné hypotézy. Je plánováno vnést do výzkumu další výběrové soubory. Je nutné podrobit experimentům výběrové soubory v dalších jazycích a především analyzovat výše zmíněným způsobem všechny již zkoumané výběrové soubory v českém jazyce pomocí přístupu III., jak již bylo provedeno s jedním z nich, a to s Bablerovým překladem do českého jazyka. Jednou z námitek může být, že poetické texty nejsou vhodný předmět pro kvantitativní analýzu. Důvody pro volbu poetických výběrových souborů byly již zmíněny několikrát. Důvodem byla unikátnost existence několika textů v různých jazycích majících stejné sémantické pozadí. Apropos, tato metodika byla aplikována ještě na jeden výběrový soubor v českém jazyce, a to na novinový článek (Nebeský, 2009). V plánu budoucích experimentů také je výše zmíněným způsobem podrobit exploraci devět překladů Poeova *The Raven* do slovenského jazyka, dostupný v (Poe, 2004). Dále je nutné podrobit experimentům jiné než poetické texty, plánovány jsou analýzy textů politologických a např. mluvené řeči. Samostatnou a velice důležitou kapitolou pro další výzkum je korektní stanovení jednotek pro tento typ kvantitativní analýzy. Je navrženo, striktně odlišovat úrovně akustické, systematické a grafické, pro každou takovou úroveň stanovit posloupnost jednotek a pomocí těchto posloupností segmentovat dané výběrové soubory a provést kvantitativní analýzu dle výše zmíněné metodiky.

5 *Havran* a teorie informace

5.1 Teorie komunikace, teorie informace a numerická estetika

Důležitým pojmem používaným v teorii informace je pojem **entropie.** Tento pojem (z řeckého entropein = odvracet) použil v roce 1865 *Rudolf Clausius* jako míru neurčitosti tepelného pohybu molekul v termodynamice. Znamená míru neurčitosti pokusu, míru neuspořádanosti systému či pro lingvistiku množství informace obsažené v jednom komunikačním signálu či znaku. Jednotkou entropie je BIT (vzniklo ze slov binary digit), která má pouze dvě podoby, 1 a 0. Je to veličina odvozená od pravděpodobnosti svých složek a určuje míru informace, jakou je systém schopen nést.[36] Entropie je tím vyšší, čím je prvek méně předvídatelný, neboli jinými slovy čím vyšší je entropie prvku, tím nižší je jeho výskyt v textu, tedy je v textu důležitější a ten by jeho vynecháním utrpěl na srozumitelnosti[37], viz (Bartók & Janoušek, 1980).

$$H = -\sum_{i=1}^{N} p_i \cdot \log_2 p_i \tag{56}$$

N	počet všech různých použitých znaků v množině
p_i	pravděpodobnost výskytu i-tého znaku
n	počet všech různých použitých znaků v množině
N_i	počet všech znaků i-tého typu[38]

Podobným způsobem je také možno zjistit *míru informace částečného i-tého znaku* ve zprávě.

$$H_i = -\log_2 p_i = \log_2 \frac{1}{p_i} = \log_2 \frac{N}{N_i} \tag{57}$$

Maximální entropii $H_{\max}$ zjišťujeme pro rovnoměrné rozložení všech znaků, tedy jestliže jsou si pravděpodobnosti jejich výskytu rovny.

$$H_{\max} = \log_2 n \tag{58}$$

Na základě znalosti entropie H a celkového počtu znaků můžeme vypočítat *informační obsah* zprávy:

---

[36] Logaritmická míra velikosti variety (jejíž jednotkou je bit) je výhodná, protože násobení lze nahradit pouhým sčítáním. Toto tvrzení je možno ilustrovat následujícím příkladem. Farmář dokáže na své farmě rozeznat osm různých druhů kuřat, nerozezná je však dle pohlaví. Jeho žena kuřata rozezná podle pohlaví, nerozlišuje však odrůdy. Dohromady jsou manželé schopni rozlišovat 2 x 8 = 16 různých „druhů" kuřat. Promluvíme-li však jazykem binární soustavy, dokáže farmář rozlišovat varietu 3 bitů a jeho žena 1 bitu, takže dohromady oba dva rozlišují varietu 3 + 1 = 4 bitů, viz (Pavlík, 2004).

[37] Kód, který má vyšší entropii, je úspornější. [1]

[38] Dle [1] stanovuje výše zmíněný vzorec při praktických aplikacích pouze odhad skutečné entropie $\hat{H}$ na základě četností jednotlivých použitých znaků a má obvykle systematickou chybu, je vychýlený.

$$I = N \cdot H \tag{59}$$

V případě, že zpráva o délce N obsahuje právě N různých znaků, pak platí, viz (Bartók & Janoušek, 1980):

$$I = N \cdot H = -N \cdot \sum_{i=1}^{N} \frac{1}{N} \cdot \log_2 \frac{1}{N} = N \cdot \log_2 N \tag{60}$$

Z předchozího vyplývá, že entropie H nemá sémantický obsah a také že nezávisí na postupnosti znaků, proto se zavádí takzvaná *střední (průměrná) entropie*, která už závisí na charakteru uspořádání, viz (Bartók & Janoušek, 1980):

$$\overline{H} = p(z_1) \cdot H_1 + p(z_2) \cdot H_2 + \ldots \tag{61}$$

Tato podmínka (závislost na charakteru uspořádání) je důležitá zejména pro hudební a textové řetězce.

Dále je možno určit vztah mezi entropií *H* a mezi maximální entropií $H_{\max}$. Tato míra se nazývá *redundance* (nadbytečnost) H, viz (Bartók & Janoušek, 1980):

$$R = \frac{H_{\max} - H}{H_{\max}} = 1 - \frac{H}{H_{\max}} \tag{62}$$

Je také možno zavést *relativní redundanci* h:

$$h = \frac{H}{H_{\max}} \tag{63}$$

Přitom platí:

$$R = 1 - h \tag{64}$$

Redundance zvyšuje nadbytečnost zprávy, tím se snižuje efektivnost přenosu v kanále, což odporuje principu ekonomie v jazyce, na druhé straně to však přispívá ke spolehlivosti přenosu zprávy. Čím vyšší je frekvence jistého prvku, tím vyšší je pravděpodobnost jeho výskytu, a tedy i jeho redundance. Tím menší je pak množství přenášené informace, nebo také míra neurčitosti a entropie. Redundance je v běžném jazyce nezbytná, aby se odstranil vliv různých poruch a šumů v komunikaci, jako jsou například nedbalá výslovnost, nepozornost, překlepy a poruchy telefonního spojení. Bez redundance by mohlo dojít k nesrozumitelnosti a snížené vnímatelnosti, připomeňme si například nutnost opakování nových a neznámých pojmů při přednáškách.[39] O něco jinou funkci může mít redundance u uměleckých děl, kde nemusí jít jen o snahu o předcházení nesrozumitelnosti, ale například o zvýšení napětí nebo vyvolání jistého pocitu.

[39] Redundanci je tedy možno zvyšovat například opakováním zprávy.

Veličina, která hodnotí míru přenosu informací v kanále, se nazývá *informační tok* $\bar{I}$, viz (Bartók & Janoušek, 1980)**:**

$$\bar{I} = \frac{I}{T} \tag{65}$$

I informační obsah v bitech

T čas v sekundách

Tento vztah udává, jaké množství informací se přenese za jednotku času v sekundách. Jelikož se přenos realizuje prostřednictvím kanálu, je také důležité zjistit jeho přenosové vlastnosti, které udává *kapacita kanálu* C_k, tj. maximální informační tok, který může kanál propustit:

$$C_k = Sup\left(\bar{I}\right) \tag{66}$$

Z předchozích tvrzení je zřejmá platnost takzvané *Shannonovy podmínky*, viz (Bartók & Janoušek, 1980)**:**

$$C_k > \bar{I} \tag{67}$$

To znamená, že kapacita kanálu musí být větší než informační tok, jinak kanál celý tok nepropustí a dojde k omezení zprávy.[40]

Další relevantní uplatnění teorie informace je v numerické estetice. Výše zmíněná entropie lze počítat s přihlédnutím k významu znaků a je tedy pak založena na vztahu pole znaků $Z = \{z_j\}$ a pole významů $Z = \{v_i\}$. Potom estetická informace odpovídá entropii, viz (Bartók & Janoušek, 1980):

$$H_v(Z) = -\sum_i p_i \sum_j p(j) \cdot \log_2 p_i(j) \tag{68}$$

$p_i(j)$ pravděpodobnost, že po odevzdání významu v_i byl použit znak z_j

Dále ještě platí $0 \le H_v(Z) \le H(Z)$, což znamená, že zavedená estetická informace je menší, než informační obsah. Je to logické, protože volnější vazby rozhodně zvyšují estetickou informaci.

Numerickou estetiku dále rozvíjel *Fred Attneave* zavedením dalších veličin, dále viz (Bartók & Janoušek, 1980). Hodnota *překvapení*:

---

[40] Je nutné si v této souvislosti uvědomit, že komunikační kanál představují mimo jiné smyslové orgány spolu s vyšším nervovým centrem. Na základě psycho-fyziologických pokusů z let 1959-62 bylo zjištěno, že vnímání je učeno krátkodobou pamětí (tj. schopností zapamatovat si odděleně uvědomění jednotlivých znaků za sebou) pohybující se v rozmezí 5 s < T < 12 s. Obvykle se však bere do úvahy tzv. prezenční čas T = 8 s. Dalšími pokusy se zjistilo rozpětí kapacity kanálu (lidských smyslů) - 12 – 25 bit.s^{-1}, jako odhad se tudíž bere $C_k = 16\text{bit.s}^{-1}$. Obsah krátkodobé paměti (tj. vnímaná informace) je pak daný jako $K_k = C_k \cdot T = 128\text{bit}$, viz (Bartók & Janoušek, 1980).

$$U = \frac{\log_2 \frac{1}{p_i}}{H} = \frac{-\log_2 p_i}{-\sum_i p_i \cdot \log_2 p_i} \tag{69}$$

To je vlastně poměr míry informace částečného znaku ku entropii celého systému. *Frank* dále doplňuje, že znaky s hodnotou překvapení U=1 se označují jako neutrální. Objekt má tuto hodnotu, jestliže jeho informace odpovídá informaci od něj očekávané. Nerovnost U<1 indukuje banálnost díla, naproti tomu je-li U>1, znak je opravdu překvapivý.

Další veličina, kterou Attneave zavedl, je *nápadnost*, viz (Bartók & Janoušek, 1980)**:**

$$a(z_k) = p_k \cdot u(z_k) = \frac{-p_k \cdot \log_2 p_k}{H} \tag{70}$$

p_k relativní početnost znaku, při dokončení četby díla, učení apod. platí, že $p_i = p_k$.

Platí zde dále, že $0 \le a(z_k) \le 1$ a $\sum_{k=1}^{r} a(z_k) = 1$.

To znamená, že vyskytuje-li se nějaká veličina velmi často a pokaždé s velkou mírou překvapení, pak je nápadná. Maximální nápadnost $a(z_k) = \max$ nastává pro $p_k = \frac{1}{e}$.[41]

Estetická entropie je vyjádřením originality, což vlastně znamená:

1. Nejen relativní, ale i absolutní hodnota estetické informace se zvyšuje při zmnožení významu znaků (což vede k abstrakci).
2. Zákon omezení množství druhů znaků určuje styl.
3. Efekt maxima určuje výraz.

Další veličinou, kterou je možné měřit, je *estetická míra*. Materiální objekt je estetický, jestliže funguje v komunikativním procesu jako přenašeč signálu a konstelace těchto signálů přitom přenáší estetickou informaci. Estetický objekt putuje mezi producentem a příjemcem a měření samotného objektu mají rozhodně co dělat s estetickou mírou.

Birkhoff definoval estetickou míru následujícími způsoby, viz (Bartók & Janoušek, 1980):

1. Estetická míra je skalár a vypovídá něco o zalíbení (pleasingness), které objekt vyvolá v příjemci.
2. Estetická míra závisí na veličinách, které jsou objektem určeny: řád a komplexnost.

[41] Experimentálně bylo ukázáno, že maximum nápadnosti se docílí při $h_i \approx 0{,}37$, což znamená při 37% výskytu. Například psychologické testy s posluchači výtvarných škol dokázaly, že dominantní barvy obrazů pokrývají 40% obrazové plochy. V tomto případě se hovoří o *efektu maxima*, což znamená, že stylistický význam mají znaky s početností 0,33< h_i <0,47, viz (Bartók & Janoušek, 1980).

$$M = f(O,C),$$

kde *C* je míra úsilí smyslových orgánů vynaložená na vnímání objektu a *O* jako odměna za vynaložené úsilí.

3. Samotnou funkci potom definoval jako závislost

$$M = f\left(\frac{O}{C}\right).$$

4. Birkhoffovská estetická míra je dána takto:

$$M = \frac{O}{C}.$$

Ovšem podle *Eysencka* je estetická míra definována pomocí poněkud odlišné závislosti:

$$M = O \cdot C$$

Pokud předpokládáme interpretaci řádu *O* jako redundance a komplexnosti *C* jako entropie informace, pak dostáváme úpravou Birkhoffovskou estetickou míru jako

$$M = \frac{1}{H} - \frac{1}{H_{\max}} \tag{71}$$

a míru Eysenckovu jako

$$M = (H_{\max} - H)\frac{H}{H_{\max}}$$ [42], (72)

dále viz (Bartók & Janoušek, 1980), (Benešová, 1999) a (Benešová, 2010).

5.2 Vyhodnocení výpočtů

Dříve, než zahájím komentář výsledků výpočtů získaných aplikací veličin zmíněných v předchozí kapitole, považuji za důležité zmínit, že při kvantitativním uchopení jakýchkoli aspektů jazyka a jeho produktů, je nutné neztratit ze zřetele významovou stránku, což není vždycky jednoduché.[43]

V této kapitole bych ráda demonstrovala výpočty entropie a ostatních veličin zmíněných v předchozí části. Jako prvotní problém se opět jeví definování jednotlivých znaků z hlediska lingvistického.[44] Na tomto místě porovnám nejprve výsledky tří možných způsobů

[42] Použitím Birkhoffovy a Eysenckovy formule k výpočtům dostáváme dramaticky, o řády rozdílné výsledky, ale poměrné porovnání estetičnosti objektů je stejné.

[43] Přenášená informace je vztahem mezi znaky zprávy a okolním světem. Mezi znaky samotnými existují strukturní vztahy. Vztahy mezi symboly a okolním světem jsou omezené jednak na vztahy mezi označením a významem a jednak mezi významem a jejich překladem. Rozlišujeme tři varianty informace, syntaktickou, sémantickou a pragmatickou. [1]

[44] Se stanovením lexikální jednotky se potýkali mnozí kvantitativní lingvisté, jako příklad mohu uvést autory frekvenčních slovníků čeština a slovenštiny J. Jelínka, J.V. Bečky a M. Těšitelové (*Frekvence slov, slovních druhů a tvarů v českém jazyce*, Praha 1961) a J. Mistríka (*Frekvencia slov v slovenčine*, Bratislava

vyhodnocení originálního Poeova textu, dále šestnáct českých překladů s jednotkami stanovenými jedním způsobem, vyhodnotím výsledky výpočtů týkajících se speciálně jednotlivých znaků originálního textu a na závěr porovnám výpočty odrážející klíčové znaky z originálního textu i překladů a ty výpočty, které se týkají vybraných korespondujících si znaků z originálního textu a překladů.

Nejprve bych se chtěla podrobněji zmínit o postupu při vyhodnocování textu. Postup demonstruji na originálním Poeově textu s jednotkami stanovenými z grafického hlediska jako slova „od mezery k mezeře". V tab. č. 22[45], příloha X. jsem abecedně seřadila všech $n = 429$ různých použitých znaků (slov) z celkových $N = \sum_{i=1}^{n} N_i = 1079$. Nerozlišovala jsem synsémantická, modální a autosémantická slova s ohledem na to, že pokud by báseň četl nebo poslouchal někdo, kdo nemluví Poeovou angličtinou, stěží by tyto kategorie z hlediska váhy jejich významu rozlišil. I pro běžného dnešního uživatele angličtiny by slova jako *thy* či *thou* mohla působit jistý problém, neboť dnes pro svou archaičnost nejsou takřka používána.

V dalších sloupcích tab. č. 22 čteme veličiny N_i – počet všech znaků *i*-tého typu, p_i – pravděpodobnost výskytu *i*-tého znaku, H_i – míra informace částečného *i*-tého znaku, U_i – míra překvapení *i*-tého znaku, A_i – míra nápadnosti *i*-tého znaku.

První počítanou veličinou byl informační obsah zprávy, viz (59):

$$I = N \cdot H = -N \cdot \sum_i p_i \log_2 p_i = N \cdot \log_2 N - \sum_i N_i \cdot \log_2 N_i$$

$$I = 8494,09\text{bit}$$

Z toho pak také snadno můžeme zjistit entropii zprávy, viz (59):

$$H = \frac{I}{N}$$
$$H = 7,87\text{bit}$$

Maximální entropie, které by bylo dosaženo při rovnoměrném rozložení všech použitých znaků, se rovná, viz(58):

$$H_{\max} = \log_2 n$$
$$H_{\max} = 8,74\text{bit}$$

Pomocí maximální entropie je možné zjistit hodnotu redundance, viz (62):

1969). Autoři českého frekvenčního slovníku stanovili slovoformu *byl bych šel* jako jeden znak uvedený pod heslem *jít*. Na druhé straně ve slovenském frekvenčním slovníku vystupuje tvar *bol by som šiel* jako čtyři různé znaky, je na ně tedy pohlíženo důsledně jako na grafické jednotky, viz (Černý, 1996).

[45] Do příloh je pro velkou obsáhlost zařazena pouze jediná tabulka ilustrující postup při výpočtech. Ty však byly provedeny u originálního anglického textu Poeova *Havrana* a u všech jeho překladů do českého jazyka z (Poe, 1985).

$$R = 1 - \frac{H}{H_{\max}}$$
$$R = 0{,}09979 = 9{,}98\%$$

Zjištěná hodnota redundance se ocitla mimo interval mezi 0,60 a 0,80, což jsou obvyklé hodnoty redundance v jednotlivých jazycích. Dokonce je míra nadbytečnosti použitých znaků v porovnání s obvyklými mezemi neporovnatelně nízká. Z toho a z ostatních výpočtů vyplývá, že velká většina v díle použitých znaků (slov) nese velké množství informací a s největší pravděpodobností by například čtení *Havrana* po telefonu mohlo způsobit příjemci velké problémy, pokud by spojení nebylo příliš kvalitní. Je ovšem nutné si uvědomit, že nejde o běžný příklad užití jazyka a jistě zde nejde jen o spolehlivost přenosu zprávy a srozumitelnost.

Pro informační estetiku je důležitá další veličina, která hodnotí míru přenosu informací, a to informační tok. Jak jsem již zmínila, pro její výpočet bereme v úvahu jako čas *T* takzvaný prezenční čas $T = 8\text{s}$, viz (65).

$$\overline{I} = \frac{I}{T}$$
$$\overline{I} = 1061{,}76\text{bit.s}^{-1}$$

Což znamená, že

$$\overline{I} > C_k \doteq 16\text{bit.s}^{-1},$$

takže je porušena Shannonova podmínka, což znamená, že informační tok básně Havran daleko přesahuje odhad kapacity kanálu, viz pozn. 35, – lidských smyslů a dílo si není možné snadno zapamatovat a snadno osvojit. Vnímatelnost a zapamatovatelnost bez opakování je ohrožena. Pro ilustraci náročnosti vnímání zprávy mající takovýto informační tok bych ráda uvedla malé vysvětlení. *Bit* je jednotka běžně používaná v informatice. Reprezentuje stav 1 nebo 0 (zapnuto / vypnuto, ano / ne), tedy jeden ze dvou znaků dvojkové soustavy. Každý znak (v tomto případě znak = písmeno, interpunkční znaménko, symbol a další speciální funkce) na klávesnici je reprezentován jistou kombinací osmi jedniček a nul. Celosvětově je tato reprezentace známa jako ASCII kód. Jedna tato osmibitová kombinace se nazývá *byte*. To znamená, že náš vypočtený informační tok můžeme pro ilustraci převést na jiné jednotky a dostaneme:

$$\overline{I} = 1061{,}76\text{bit.s}^{-1} = 132{,}72\text{byte.s}^{-1}$$

Jinými slovy, abychom mohli plně vnímat a zapamatovat si tuto báseň, museli bychom být schopni zachytit přibližně 133 bytů neboli 133 znaků, v našem případě písmen (při poslechu fonémů) za jedinou sekundu.

Poslední veličinou týkající se celého objektu je estetická míra, pojatá však dvěma různými autory – Birkhoffem a Eysenckem, viz (71) a (72).

		Havran
estetická míra podle Birkhoffa	$M = \frac{1}{H} - \frac{1}{H_{\max}}$	$M = 0{,}013$
estetická míra podle Eysencka	$M = (H_{\max} - H)\frac{H}{H_{\max}}$	$M = 0{,}79$

Tab. č. 12: Výpočty estetické míry dle Birkhoffa a Eysencka

Hodnoty výpočtů (viz. tab. č. 12) obou autorů se sice značně liší, avšak poměrem vyjadřují obdobné výsledky.

5.3 Porovnání různých způsobů vyhodnocení originálního textu *Raven*

Již několikrát jsem zmínila nesnáze, s nimiž se kvantitativní lingvista může setkat na počátku své snahy o vyhodnocení libovolného objektu, výběrového souboru. Jedním z prvních kroků v algoritmu by mělo být stanovení jednotek, s nimiž se bude dále operovat. Abych ilustrovala rozdíly, které mohou nastat ve výsledcích při různých způsobech stanovování jednotek, porovnám tři různá kvantitativní zpracování originální básně *Raven*, tj. různé výsledky při různém stanovení jednotek.

V prvním, již výše zmíněném způsobu, volím znaky stejným způsobem, jako bylo např. učiněno J. Mistríkem ve slovenském frekvenčním slovníku, tedy znakem budiž slovo z grafického náhledu „od mezery k mezeře“. Tato volba příliš neodráží přenášenou informaci ani v její podobě gramatické, ne vždy v podobě sémantické a ani v podobě pragmatické. Situace je o to složitější, že anglický jazyk je primárně analytický jazyk.

Druhý způsob respektuje cestu vyznačenou například kolektivem autorů českého frekvenčního slovníku *Frekvence slov, slovních druhů a tvarů v českém jazyce* a za jednotku pojímá celou slovoformu, například *is sitting*.

Třetí, poslední způsob akcentuje výpočet entropie na základě četnosti výskytu jednotlivých lexémů, například slovoformě *is sitting* byl přiřazen lexém *to sit*, jehož četnost a z ní vyplývající konsekvence byly následně vypočítány.

	Raven (znak - slovo "mezera-mezera")	*Raven* (znak - slovoforma)	*Raven* (znak - lexém)
různé použité znaky	429	447	412
celkový počet znaků	1079	1060	1060
informační obsah (bit)	8494,09	8395,49	8245,23
entropie (bit)	7,87	7,92	7,78
maximální entropie (bit)	8,74	8,8	8,69
redundance (%)	9,98	10,04	10,45
informační tok (bit.s^{-1})	1061,76	1049,44	1030,65
estet. míra (Birkhoff)	0,013	0,01	0,01
estet. míra (Eysenck)	0,79	0,8	0,81

Tab. č. 13: Porovnání výsledků zpracování originálního textu *Raven* různými způsoby

Z výsledků v tab. č. 13 je patrné, že množství přenášené informace u tří různých kvantitativních zpracování originálního Poeova textu je srovnatelné, odchylka u entropie se pohybuje v desetinách, podobně u redundance. Informační tok ve všech třech případech se liší v řádu jednotek a zároveň ve všech třech případech velice překračuje kapacitu lidských smyslů, střední hodnota 16 bit.s^{-1}. Obě estetické míry ve všech třech případech při daném zaokrouhlení vycházejí takřka totožně.

Při bližším pohledu na tři zpracování básně *Raven* je patrné, že nejvyšší informační obsah byl naměřen při zpracování textu s jednotkami skupinami písmen mezi dvěma mezerami. Na druhé straně nejvyšší entropie nese text s jednotkami stanovenými jako slovoformy. Nejvyšší redundanci má z pochopitelných důvodů zpracování s jednotkami lexémy, neboť se takto procesem zvaným lematizace několik různých slovoforem sdruží do jedné množiny jako jeden lexém a ten pak má samozřejmě vyšší frekvenci, opakuje se a dochází k redundanci. Estetické míry, jak již bylo řečeno, jsou ve všech třech vyhodnoceních takřka stejné.

5.4 Porovnání výsledků výpočtů šestnácti českých překladů básně *Raven*

V této kapitole přináším porovnání příslušných výpočtů týkajících se šestnácti českých překladů Poeovy originální básně *Raven*, viz (Poe, 1985). Pro porovnání připojuji také výsledky týkající se originálního textu. Všech šestnáct překladů bylo vyhodnoceno druhým a třetím způsobem popsanými v předchozí části, tedy nejprve slovoforma a poté lexém byly definovány za jednotku, viz obr. č. 14 a 15.

		informační obsah zprávy	entropie	max. entropie pro rovn. rozložení všech použitých znaků	redundance	Informační tok
		I	*H*	*Hmax*	*R*	$\bar{I}$
Poe	slovoformy	8395,493790	7,920277	8,804131021	0,10039081	1049,436724
	lexémy	8245,235907	7,778524	8,686500527	0,10452726	1030,654488
Šembera	slovoformy	7353,925085	8,501647	9,016808288	0,05713339	919,2406356
	lexémy	6940,503920	8,023704	8,73470962	0,08140003	867,56299
Vrchlický	slovoformy	7163,552023	8,368636	8,982993575	0,06839124	895,4440028
	lexémy	6861,639752	8,015934	8,714245518	0,08013444	857,704969
Mužík	slovoformy	7434,012736	8,634161	9,063395081	0,04735907	929,251592
	lexémy	7164,671384	8,321337	8,839203788	0,05858746	895,583923
Lutinov	slovoformy	7097,021732	8,654905	9,047123912	0,04335293	887,1277165
	lexémy	6874,866391	8,383983	8,839203788	0,05150016	859,3582989
Nezval	slovoformy	6971,270739	8,269598	8,903881846	0,07123683	871,4088424
	lexémy	6684,701729	7,929658	8,696967526	0,08822725	835,5877161
Babler	slovoformy	7190,908109	8,381012	8,918863237	0,06030493	898,8635136
	lexémy	6851,478406	7,985406	8,607330314	0,07225519	856,4348008
Taufer	slovoformy	7144,522079	8,485181	8,971543554	0,05421173	893,0652599
	lexémy	6811,698132	8,089903	8,693486957	0,06942947	851,4622665

		informační obsah zprávy	entropie	max. entropie pro rovn. rozložení všech použitých znaků	redundance	Informační tok
Stoklas	slovoformy	7388,479000	8,492505	8,991522000	0,05550000	923,5599000
	lexémy	7039,005000	8,09081	8,734710000	0,07370000	879,8756000
Wagnerová	slovoformy	7720,183914	8,355177	8,982993575	0,06988942	965,0229892
	lexémy	7343,285095	7,947278	8,721099189	0,08872975	917,9106369
Havel	slovoformy	7136,184379	8,577145	8,974414590	0,04426694	892,0230473
	lexémy	6789,316004	8,160236	8,703903573	0,06246255	848,6645005
Čapek	slovoformy	6285,764053	8,482812	8,888743249	0,04566796	785,7205067
	lexémy	6019,101385	8,122944	8,629356620	0,05868488	752,3876732
Resler	slovoformy	7427,968548	8,54772	9,057991723	0,05633389	928,4960685
	lexémy	7129,849397	8,20466	8,804131021	0,06808976	891,2311746
Černý	slovoformy	6804,129993	8,494544	8,962896005	0,05225451	850,5162491
	lexémy	6574,190746	8,207479	8,778077130	0,06500262	821,7738433
Slavík	slovoformy	7018,363473	8,295938	8,915879379	0,06953229	877,2954342
	lexémy	6746,102847	7,974117	8,682994584	0,08163978	843,2628559
Kadlec	slovoformy	7553,461321	8,54464	8,994353437	0,04999958	944,1826651
	lexémy	7090,010903	8,020374	8,658211483	0,07366847	886,2513629
Bejblík	slovoformy	7178,839941	8,396304	8,971543554	0,06411823	897,3549927
	lexémy	6854,642544	8,017126	8,717676423	0,08035979	856,830318

Tab. č. 14: Porovnání výsledků výpočtů týkajících se šestnácti českých překladů Poeovy básně *Raven*

		estetická míra - Birkhoff	estetická míra - Eysenck
		M_B	M_E
Poe	slovoformy	0,012675164	0,795123054
	lexémy	0,013437929	0,813067836
Šembera	slovoformy	0,006720272	0,48572791
	lexémy	0,010144945	0,65312978
Vrchlický	slovoformy	0,008172328	0,572341329
	lexémy	0,009996894	0,642352445
Mužík	slovoformy	0,00548508	0,408905828
	lexémy	0,007040631	0,487526038
Lutinov	slovoformy	0,00500906	0,375215501
	lexémy	0,006142684	0,431776463
Nezval	slovoformy	0,008614304	0,58909989
	lexémy	0,011126236	0,699611883
Babler	slovoformy	0,007195424	0,505416359
	lexémy	0,009048405	0,576986994
Taufer	slovoformy	0,00638899	0,459996359
	lexémy	0,008582238	0,561677658

		estetická míra - Birkhoff	estetická míra - Eysenck
Stoklas	slovoformy	0,006535000	0,471322000
	lexémy	0,009111000	0,596433000
Wagnerová	slovoformy	0,008364803	0,583938473
	lexémy	0,011164797	0,705160008
Havel	slovoformy	0,005161035	0,37968398
	lexémy	0,007654503	0,509709104
Čapek	slovoformy	0,005383587	0,387392732
	lexémy	0,007224583	0,476694006
Resler	slovoformy	0,006590516	0,481526277
	lexémy	0,008298913	0,558653336
Černý	slovoformy	0,006151537	0,443878208
	lexémy	0,007919926	0,533507675
Slavík	slovoformy	0,008381486	0,57683552
	lexémy	0,010238096	0,651005122
Kadlec	slovoformy	0,005851573	0,427228424
	lexémy	0,009185166	0,590848673
Bejblík	slovoformy	0,007636483	0,538356186
	lexémy	0,010023516	0,644254539

Tab. č. 15: Porovnání výsledků výpočtů týkajících se šestnácti českých překladů Poeovy básně *Raven* – estetické míry

Nejvyšší informační obsah nese překlad Dagmar Wagnerové vyhodnocený první metodou (7 720,18 bit) a na druhou stranu nejnižší informační obsah najdeme u Čapkova překladu hodnoceného druhou metodou (6 019,10 bit). Je nutné podotknout, že originální text vyhodnocený oběma metodami má zdaleka nejvyšší informační obsah (8 395,49 a 8 245,24 bit). Nejvyšší míra entropie byla zjištěna u Lutinovova překladu vyhodnoceného první metodou (8,65 bit) a nejnižší u překladu Nezvalova vyhodnoceného druhou metodou (7,92 bit). Opět pomíjíme vyhodnocení originálního textu druhou metodou, jež neslo entropii nejnižší (7,78 bit). K nejvyšší redundanci došlo v překladu u Wagnerové vyhodnoceném první metodou (6,99%) a k nejnižší U Lutinova druhou metodou vyhodnoceného (4,34%). Zcela nejvyšší redundanci dosáhl ale originální text vyhodnocený druhou metodou (10,45%). Informační tok proto ve všech případech mnohonásobně přesáhl kapacitu lidských smyslů, nejvyšší byl u Wagnerové kvantifikované prvním způsobem (965,02 bit.s^{-1}), pokud opět nepočítáme originál vyhodnocený oběma způsoby (1 049,34 a 1 030,65 bit.s^{-1}), nejnižší informační tok byl zjištěn v Čapkově překladu hodnoceném oběma způsoby (785,72 a 752,39 bit.s^{-1}). Nejvyšší estetické míry bylo dosaženo mimo originálu (0,013 dle Birkhoffa a 0,8 dle Eysencka oběma způsoby) u Wagnerové vyhodnocené druhým způsobem (0,011 dle Birkhoffa a 0,7 dle Eysencka) a nejnižší u Lutinova vyhodnoceného prvním způsobem (0,005 dle Birkhoffa a 0,38 dle Eysencka).

5.5 Vyhodnocení výpočtů týkající se speciálně jednotlivých znaků

V předchozích částech této kapitoly jsem se pokusila komentovat výpočty týkající se celého objektu – básně *Raven* či českých překladů. Nyní bych se ráda věnovala detailněji

speciálně několika příkladům použitých znaků v originálním textu – slov, která vykazují charakteristické vlastnosti. Slova jsem si seřadila podle rostoucí frekvence (viz tab. č. 16) a získala tak frekvenční seznam, ve kterém je možné, podobně jako ve frekvenčních slovnících rozlišit tři pásma slov. Z prvního pásma slov s nejvyšší a vyšší frekvencí jsem si zvolila slovo s absolutně nejvyšší četností – anglický určitý člen *the*. Ze skupiny slov s nízkou a nejnižší četností jsem vybrala jedno ze slov s četností $N=1$ - adjektivum *black* a z prostřední skupiny s frekvencí střední jsem vybrala slovo, jenž sám Poe označil pro svou báseň za extrémně důležité – adverbium *nevermore*.

	THE	**NEVERMORE**	**BLACK**
N_i	57	11	1
p_i	5,283%	1,019%	0,093%
H_i *(bit)*	4,24259	6,61605	10,0755
U_i	0,538934	0,840433	1,279882
A_i	2,847%	0,857%	0,119%

Tab. č. 16: Porovnání výsledků výpočtů u parciálních znaků vybraných z originálního textu

Zcela pochopitelným a předvídatelným závěrem vyplývajícím z tabulky je, že největší množství informace nese znak (slovo *black*), který se v básni objevuje pouze jedenkrát. Tento znak je zároveň nejvíce překvapující. Je ale nutné vzít v úvahu celkový počet znaků (*Havran: N*=1079). Závěrem tedy je, že hodnota překvapení je o to větší, čím větší je celkový počet použitých znaků. Fakt, že hodnota překvapení je větší než 1, indukuje opravdovou překvapivost znaku. Ke slovům s nejmenší četností v básni je třeba ještě podotknout, že se většinou jedná o slova autosémantická. Důležitost těchto slov je ohromná a ztrátou většiny z nich by mohlo dojít k ohromnému poškození vnímatelnosti a srozumitelnosti.

Na druhé straně pokud pohlédneme na nejčetnější slovo v básni – anglický určitý člen *the*, vidíme, že si ve srovnání s ostatními nese nejmenší množství informace, není také slovem plnovýznamovým, kategorii členu ostatně mnohé jazyky postrádají. V angličtině nás však jeho existence nepřekvapí, což ostatně dokládá nízká míra překvapení. To, že je tato míra menší než jedna, vyjadřuje banálnost znaku. Pokud bychom tedy při poslechu básně po telefonu některý z určitých členů nezachytili, s největší pravděpodobností by se s vnímáním a srozumitelností celé básně příliš nestalo nebo bychom si ho v lepším případě s jistou úrovní znalosti angličtiny byli schopni sami domyslet. Tyto závěry vyplývající z tabulky pro slovo THE v angličtině ale samozřejmě neznamenají banálnost jeho existence v jazyce, viz. např. rozdíl mezi

the brother of mine (ten konkrétní nebo ten jediný)

a brother of mine (jeden ze všech, neznámo který)

* brother of mine (který vůbec? o kom se to mluví?).

Již výše jsem upozornila na nutnost neztratit ze zřetele stránku významovou. A jak bude patrné také dále, opakování znaků jen nezvyšuje nutně nadbytečnost zprávy, čímž by s výjimkou situací jako například opakování při výuce, komunikace mezi mluvčími různých mateřských jazyků či pokročilostí hrubě porušuje princip ekonomie v jazyce.

Dalším aspektem, který je nutné v souvislosti s tímto znakem – slovem *the* – zmínit, je znovu problém stanovení jednotek. Člen v anglickém jazyce plní sémanticko-gramatickou funkci a „syntakticky má funkci determinátoru, tj. nesamostatného větného členu v rámci větného členu realizovaného substantivem, v němž zpravidla tvoří první složku, tj. předchází před premodifikací“, viz (Dušková, 1994). Z toho vyplývá, že by též přicházelo v úvahu počítat jako jeden znak člen dohromady se substantivem, které rozvíjí, jak také navrhuje Hřebíček[46], viz (Hřebíček, 1997). Zde se však chci zaměřit na funkci, použití a entropii členu *the* jako takového, proto jej záměrně ponechávám jako samostatný znak.

Jako poslední znak – slovo - k rozboru jsem zvolila slovo *nevermore*, jehož četnost je 11. Patří tedy k nevelké skupině četnějších slov, ve které se, jak už jsem zmínila, vyskytují slova mající spíše význam gramatický nebo provazují jednotlivé části básně – členy, spojky, předložky, příslovce a zájmena. Výjimkami s četností nad 5 jsou pouze *bird* (7), *bust* (6), *door* (14), *chamber* (11), *Lenore* (8), *raven* (6), *said* (6) a *soul* (6). A pak také zmíněné slovo *nevermore*. Míra překvapení u slov s touto četností je opět menší než jedna, což by mělo vyjadřovat jejich banálnost. Nicméně v tomto případě bych tento jev opět nenazývala banálností, jako spíše snahou autora navodit a udržet atmosféru, kterou navodil. Stěží při poslechu poslední sloky zapomeneme, že démon, který hlavního hrdinu straší, má podobu havrana, i kdybychom slovo *raven* v této chvíli přeslechli. A zřejmě budeme i na konci poslední sloky očekávat zakrákání *nevermore*. Nepřekvapí nás to, ale jistě má opakování slov v básni jiný význam. O jeho důležitosti píše sám autor ve své *The Philosophy of Composition,* kde jej nazývá refrénem a vysvětluje, jak jej hledal a jaká pro něj stanovil kriteria, včetně těch fonetických, viz (Poe, 1985). O významu opakování těchto slov bude ještě zmínka.

5.6 Porovnání kvantifikací refrénů

„… hledal jsem nějakou uměleckou dráždivost, která by mi posloužila při skládání básně – nějaký čep, na němž by se mohla celá stavba otáčet,“ viz (Poe, 1985). Těmito slovy uvedl Edgar Allan Poe důvod použití refrénu ve své básni *Raven* a dále svou myšlenku rozvíjí s tím, že by refrén měl být krátký úderný a na konci každé sloky se opakovat v jiném kontextu. Dále požadoval, aby refrén obsahoval konsonantu *r* a vokál *o* z formálního důvodu navýšení zvučnosti a sémanticky aby odrážel motiv smutku. Výsledkem bylo slovo *nevermore*, viz (Poe, 1985).

A před úkolem splnit nejen všechny tyto požadavky týkající se jen samotného refrénu stál též každý překladatel, který se rozhodl pustit do nesnadného úkolu najít důstojný ekvivalent originální Poeovy básně v jiném jazyce. Každý překladatel se s tímto těžkým úkolem vypořádal jiným způsobem podnícen odlišnou motivací:

Poe – *nevermore*
Šembera – *nikdy víc*
Vrchlický – *nikdy víc*

[46] Dle Duškové, viz (Dušková, 1994), se člen dále od premodifikátoru liší tím, že příslušný větný člen v angličtině provází obligatorně. Pokud tedy budeme počítat se členy jako znaky nebo částmi znaků, je třeba zvážit nulovou variantu členu neurčitého a její započítání jako znak či součást znaku.

Mužík – *nadarmo*
Lutinov – *nevermore*
Nezval – *už víckrát ne*
Babler – *marný blud*
Taufer – *nikdy již*
Stoklas – *nikterak*
Wagnerová – *vrať mi čas, vrátit čas, nenavrátí čas, zvrátit čas*
Havel – *ni jedenkrát*
Čapek – *stokrát ztraceno*
Resler – *marnost-zmar*
Černý – *nikdá ne*
Slavík – *víckrát ne, nikdy ne*
Kadlec – *nikdykrát*
Bejblík – *marno vše*

V tab. č. 17 přináším výsledky spojené s klíčovými slovy objevujícími se právě v refrénech originálu a jeho překladech.

		N_i	p_i	H_i	$N_i.log_2N_i$	U_i	A_i	$pi.H_i$
	slovoformy	četnost každého lexému						
Poe	nevermore	11	0,010377	6,590417	38,05374781	0,832094	0,008635	0,068391
Šembera	víc	17	0,019653	5,669093	69,4868683	0,666823	0,013105	0,111416
	nikdy	16	0,018497	5,756556	64	0,677111	0,012525	0,10648
Vrchlický	nikdy	11	0,01285	6,282035	38,05374781	0,750664	0,009646	0,080727
	víc	27	0,031542	4,986579	128,3819626	0,595865	0,018795	0,157287
Mužík	nadarmo	18	0,020906	5,579944	75,05865003	0,646264	0,013511	0,116654
Lutinov	Nevermore	11	0,013415	6,220048	38,05374781	0,718673	0,009641	0,08344
Nezval	už	16	0,01898	5,719389	64	0,691616	0,013127	0,108553
	víckrát	12	0,014235	6,134426	43,01955001	0,741805	0,010559	0,087323
	ne	20	0,023725	5,397461	86,4385619	0,652687	0,015485	0,128054
Babler	marný	13	0,015152	6,044394	48,10571634	0,721201	0,010927	0,091582
	blud	16	0,018648	5,744834	64	0,685458	0,012782	0,10713
Taufer	nikdy	13	0,015439	6,017237	48,10571634	0,709147	0,010949	0,092903
	již	18	0,021378	5,547751	75,05865003	0,653817	0,013977	0,118598
Stoklas	nikterak	18	0,02069	5,594947	75,05865003	0,65881	0,013631	0,115758
Wagnerová	vrať	2	0,002165	8,851749	2	1,059433	0,002293	0,01916
	mi	14	0,015152	6,044394	53,30296891	0,723431	0,010961	0,091582
	vrátit	7	0,007576	7,044394	19,65148445	0,843117	0,006387	0,053367
	nenavrátí	2	0,002165	8,851749	2	1,059433	0,002293	0,01916
	zvrátit	1	0,001082	9,851749	0	1,179119	0,001276	0,010662
	čas	18	0,019481	5,681824	75,05865003	0,680036	0,013247	0,110685
Havel	ni	12	0,014423	6,115477	43,01955001	0,712997	0,010284	0,088204
	jedenkrát	9	0,010817	6,530515	28,52932501	0,761386	0,008236	0,070643

		N_i	p_i	H_i	$N_i.log_2N_i$	U_i	A_i	$pi.H_i$
	slovoformy	četnost každého lexému						
Čapek	stokrát	11	0,014845	6,073898	38,05374781	0,716024	0,010629	0,090166
	ztraceno	15	0,020243	5,626439	58,60335893	0,663275	0,013427	0,113896
Resler	marnost	12	0,013809	6,17825	43,01955001	0,722795	0,009981	0,085315
	zmar	21	0,024166	5,370895	92,23866588	0,628342	0,015184	0,129791
Černý	nikdá	11	0,013733	6,186227	38,05374781	0,728259	0,010001	0,084954
	ne	19	0,02372	5,397731	80,71062276	0,635435	0,015073	0,128036
Slavík	víckrát	9	0,010638	6,554589	28,52932501	0,790096	0,008405	0,06973
	ne	19	0,022459	5,476586	80,71062276	0,660153	0,014826	0,122997
	nikdy	4	0,004728	7,724514	8	0,93112	0,004402	0,036523
	víc	16	0,018913	5,724514	64	0,690038	0,01305	0,108265
Kadlec	nikdykrát	12	0,013575	6,20294	43,01955001	0,725945	0,009854	0,084203
Bejblík	marno	11	0,012865	6,280349	38,05374781	0,74799	0,009623	0,0808
	vše	16	0,018713	5,739781	64	0,683608	0,012793	0,107411

Tab. č. 17: Kvantitativní vyhodnocení nejdůležitějších slov objevujících se refrénech originálního textu a jeho překladů

Frekvence velké většiny slov, které se vyskytly v refrénech, je deset a více. Slova s frekvencí nižší se vyskytla v překladech těch autorů, kteří nedodrželi striktně kompaktní překlad Poeova *nevermore* za každou z posledních jedenácti slok v básni. Je ale pochopitelné, že používat na závěr jedenácti slok naprosto stejný tvar závěrečného slova se může jevit v češtině jako flexivním jazyce složitější než v jazyce anglickém, když pomineme obtížnost tohoto úkolu jako takového. Maximální frekvence refrénů nebo jejich částí by však neměla přesáhnout $N = 18$[47], neboť v originální Poeově básni a všech jejích šestnácti zkoumaných překladech je právě osmnáct strof. Pokud je frekvence slov vyšší než osmnáct, pak se musela v textu vyskytnout víckrát než jen v refrénu, čímž samozřejmě míra překvapení jimi vyvolaná klesá.

Drtivá většina zmíněných slov použitých v refrénech zkoumaných básní se jeví jako nepřekvapivá až banální. Interpretace této veličiny rozporující banálnost již byla uvedena dříve, opakování slov má svůj význam zmíněný mimochodem ve *Filozofii básnické skladby*, viz (Poe, 1985). Jediná tři slova, jejichž míra překvapení je větší než jedna, a jsou tedy překvapivá, jsou *vrať, nenavrátí* a *zvrátit* v překladu Dagmar Wagnerové. Důvodem je, že právě ve Wagnerové překladu byl refrén nejvariabilnější a slova v něm použitá se nejvíce měnila, čtenář či posluchač tedy má nejmenší možnost odhadnout zakončení každé strofy.

[47] Pouze Mužíkovi a Stoklasovi se podařilo přeložit Poeovo *nevermore* jednoslovně a navíc, na rozdíl od Poea, dokázali všech osmnáct strof tímto refrénem i uzavřít.

5.7 Porovnání výsledků výpočtů reflektujících vybrané korespondující si slova z originálního textu i překladů

Závěrečná tab. č. 18, která si zaslouží komentáře, obsahuje data nejdůležitějších vybraných tří slov stejných pro originál i všechny překlady. Tři zvolená slova jsou *havran, pták* a jméno zemřelé milenky[48]. Nutno poznamenat, že na rozdíl od předchozích tabulek byly výsledky získány kvantifikací při výběru jednotek třetí metodou, a to definování slova jako lexému. Důvodem je, že nás zajímá výskyt všech slovoforem, podmnožin množiny lexém.

		N_i	p_i	H_i	$N_i.log_2N_i$	U_i	A_i	$pi.H_i$
	lexémy	četnost každého lexému						
Poe	bird	10	0,009434	6,72792	33,219281	0,864935	0,00816	0,063471
	raven	10	0,009434	6,72792	33,219281	0,864935	0,00816	0,063471
	Lenore	8	0,007547	7,049849	24	0,906322	0,00684	0,053206
Šembera	havran	17	0,019653	5,669093	69,486868	0,706543	0,013886	0,111416
	pták	3	0,003468	8,171594	4,7548875	1,018432	0,003532	0,028341
	Leonora	2	0,002312	8,756556	2	1,091336	0,002523	0,020246
Vrchlický	pták	10	0,011682	6,419539	33,219281	0,800847	0,009356	0,074995
	havran	8	0,009346	6,741467	24	0,841008	0,00786	0,063004
	Lenora	3	0,003505	8,156504	4,7548875	1,017536	0,003566	0,028586
	Leonora	3	0,003505	8,156504	4,7548875	1,017536	0,003566	0,028586
Mužík	havran	11	0,012776	6,290438	38,053748	0,755941	0,009658	0,080366
	pták	9	0,010453	6,579944	28,529325	0,790732	0,008265	0,06878
	Lenora	1	0,001161	9,749869	0	1,171671	0,001361	0,011324
	Leonora	1	0,001161	9,749869	0	1,171671	0,001361	0,011324
Lutinov	pták	9	0,010976	6,509555	28,529325	0,776427	0,008522	0,071446
	havran	8	0,009756	6,67948	24	0,796695	0,007773	0,065166
	Lenor	8	0,009756	6,67948	24	0,796695	0,007773	0,065166
Nezval	havran	13	0,015421	6,018949	48,105716	0,759043	0,122284	0,092819
	pták	7	0,008304	6,912034	19,651484	0,871669	0,065845	0,057395
	Lenora	7	0,008304	6,912034	19,651484	0,871669	0,065845	0,057395
Babler	pták	10	0,011655	6,422906	33,219281	0,804331	0,009374	0,074859
	havran	8	0,009324	6,744834	24	0,844645	0,007875	0,062889
	Lenora	7	0,008159	6,937479	19,651484	0,868770	0,007088	0,056599
Taufer	havran	10	0,011876	6,395748	33,219281	0,790584	0,009389	0,075959
	pták	10	0,011876	6,395748	33,219281	0,790584	0,009389	0,075959
	Lenora	4	0,004751	7,717676	8	0,953989	0,004532	0,036664
Stoklas	pták	14	0,016092	5,957517	53,302969	0,736331	0,011849	0,095868
	havran	10	0,011494	6,442943	33,219281	0,796329	0,009153	0,074057
	Lora	8	0,009195	6,764872	24	0,836118	0,007688	0,062206
Wagnerová	havran	6	0,006494	7,266787	15,509775	0,914374	0,005937	0,047187

[48] V originále je to *Lenore*, ve zkoumaných překladech potom *Lenora, Leonora, Lenor, Lora, Jarmila, Elena* a *Tereza.* [P1]

		N_i	p_i	H_i	$N_i.log_2N_i$	U_i	A_i	$pi.H_i$
	lexémy	četnost každého lexému						
	Jarmila	5	0,005411	7,529821	11,60964	0,947472	0,005127	0,040746
Havel	havran	11	0,013221	6,241008	38,053748	0,764807	0,010112	0,082513
	pták	9	0,010817	6,530515	28,529325	0,800285	0,008657	0,070643
	Leonóra	5	0,00601	7,378512	11,60964	0,904203	0,005434	0,044342
Čapek	pták	8	0,010796	6,53333	24	0,804306	0,008683	0,070535
	Elena	8	0,010796	6,53333	24	0,804306	0,008683	0,070535
	havran	7	0,009447	6,725975	19,651484	0,828022	0,007822	0,063538
Resler	havran	10	0,011507	6,441284	33,219281	0,785076	0,009034	0,074123
	pták	6	0,006904	7,17825	15,509775	0,874899	0,006041	0,049562
	Lenora	4	0,004603	7,763212	8	0,946196	0,004355	0,035734
Černý	pták	8	0,009988	6,645658	24	0,809708	0,008087	0,066374
	havran	8	0,009988	6,645658	24	0,809708	0,008087	0,066374
	Lenora	5	0,006242	7,32373	11,60964	0,892324	0,00557	0,045716
Slavík	havran	9	0,010638	6,554589	28,529325	0,821983	0,008745	0,06973
	pták	6	0,007092	7,139551	15,509775	0,895341	0,00635	0,050635
	Leonora	5	0,00591	7,402586	11,60964	0,928327	0,005487	0,043751
Kadlec	pták	12	0,013575	6,20294	43,01955	0,773398	0,010499	0,084203
	havran	6	0,006787	7,20294	15,509775	0,898080	0,006096	0,048889
	Lenora	4	0,004525	7,787903	8	0,971015	0,004394	0,035239
Bejblík	havran	12	0,014035	6,154818	43,01955	0,767709	0,010775	0,086383
	pták	9	0,010526	6,569856	28,529325	0,819478	0,008626	0,069156
	Tereza	7	0,008187	6,932426	19,651484	0,864702	0,007079	0,056757

Tab. č. 18: Kvantifikace tří stejných klíčových slov z originálního textu a jeho překladů

Slovo *bird* bylo v Poeově originále použito desetkrát, tato frekvence se udržela v pouhých třech překladech (Vrchlický, Babler, Taufer). V jednom případě byly tvary slova *pták* použity vícekrát než desetkrát ($N=12$, Kadlec) a ve dvanácti zbylých překladech méněkrát než v originále. V Šemberově překladu je dokonce frekvence tohoto slova pouze $N=3$, takže se zvyšuje i míra překvapení, kterou nese, a stává se překvapivým slovem.

Slovo *raven* Poe ve své básni použil také desetkrát a frekvenci deset si zachovalo také ve třech ze zkoumaných překladů (Taufer, Stoklas, Resler). V pěti překladech jejich autoři frekvenci navýšili a v ostatních osmi překladech ji snížili. Nejnižší dosažená četnost byla $N=6$ a tudíž ani míra překvapení neklesla pod jedna a slovo *havran* je tedy ve všech případech nepřekvapivé.

Jméno milenky Poe zmínil celkem osmkrát. Osmkrát se pak objevilo ve třech překladech (Lutinov, Stoklas, Čapek), ve zbývajících všech třinácti překladech se pak objevilo méně než osmkrát. Ve Vrchlického překladu se objevilo šestkrát, avšak kontinuita je roztříštěna, neboť třikrát Vrchlický použil variantu *Lenora* a třikrát *Leonora*, tudíž míra překvapení obou variant narostla a staly se tak překvapivými. K obdobné situaci došlo i v Mužíkově překladu, kde navíc obě varianty milenčina jména nesou nejvyšší entropii ze všech tří výskytů všech tří vybraných slov.

6. Závěr

Tato práce si klade několik nemalých cílů. Pokud je začnu jmenovat chronologicky, pak prvním, byť ne primárním, bylo ukázat, že historie využívání matematických metod v lingvistice není krátká. Matematické metody byly a jsou v lingvistice nejen používány, ale daly též vzniknout specifickému jazykovědnému odvětví, tím je matematická lingvistika. Byť byl původní účel využívání matematiky pro jazykozpytné účely jiný, dnes je nutné přijmout pomocnou ruku matematiky obzvláště pro posuzování hypotéz. Jinými slovy je třeba precizně stanovit hypotézu, „přeložit" ji do jazyka matematiky, posoudit a vhodnými metodami vyřešit tento matematický model a závěry zpět „přeložit" do jazyka lingvistiky. Není možné říci, že by jeden či druhý krok byl jednodušší či složitější než jiný, každý z nich je nutné precizovat.

V historii lingvistiky bylo vysloveno několik hypotéz o struktuře jazyka, principu linearity, souvislosti jednotlivých jazykových hladin mezi sebou vzájemně, o existenci nadvětných struktur, eventuelně o fraktálních vlastnostech jazyka. Je tedy nutné tyto hypotézy studovat a zvážit je pomocí matematických nástrojů. Ruku v ruce s tímto záměrem jde snaha o vybudování snadno uchopitelného algoritmu kvantifikace textového výběrového souboru, který by, ač rigorózně stanoven, co nejjednodušší způsobem vedl uživatele s libovolným vzděláním skrze spleť matematična. I zde, jak je ostatně při jakémkoli výzkumu poměrně běžné, vede každý krok k otázkám spíše než k odpovědím.

V prvním kroku zmíněného algoritmu jsem diskutovala volbu materiálu pro analýzu. Potvrdilo se, že je prakticky nemožné analyzovat celou populaci. Tento fakt ale znamená, že je nutné pečlivě zvolit výběrový soubor/výběrové soubory a volbu zdůvodnit. V případě tohoto experimentu byl zvolen text básně Edgara Allena Poea *The Raven*, což je sice text poetický, tedy zdánlivě nejméně vhodný pro kvantitativní analýzu, ale na druhé straně nabízí díky existenci mnoha překladů do různých jazyků jedinečnou možnost porovnat výstupy v rámci jednoho jazyka i v rámci jazyků typologicky podobých či naopak velice různých. Dokonce bylo možné porovnat překlady jednoho autora do dvou různých jazyků. Tento fakt je z výstupů experimentu velice dobře patrný. Pro porovnání byl připojen ještě jeden „nepoetický" výběrový soubor, a to text novinového článku.

Druhým krokem algoritmu je vhodné stanovení jednotek pro kvantifikaci. Byly testovány čtyři přístupy pro stanovení jednotek ‚slovo'. Jako nejefektivnější a nejlepší výsledky poskytující přístup se experimentálně prokázal ten, ve kterém je slovo chápáno jako analytická slovoforma, ke které je přiřazena předložka.

Třetím krokem algoritmu je verifikace reprezentativnosti výběrového souboru, tedy faktu, že vzorek postačujícím způsobem reprezentuje svou populaci.

Exaktní aparát, který byl účelově zvolen pro tento experiment, je Menzerath-Altmannův zákon, který byl mimo obou v názvu připomenutých lingvistů podrobně prozkoumán orientalistou Luďkem Hřebíčkem. Právě on upozornil na platnost tohoto zákona na všech jazykových hladinách, což jej zároveň vedlo k nadefinování nadvětných struktur. Povšiml si též zásadní souvislosti mezi tímto zákonem a jednou z vlastností fraktálních objektů. Tato souvislost dovolila vyslovit precizní požadavky pro existenci jazykového fraktálu. Tyto vlastnosti formalizoval Jan Andres. Bylo však nutné hypotézy testovat. Na materiálu zvolených

výběrových souborů jsem provedla experimenty a testovala též vybudování metodologické základny.

Ve čtvrtém kroku algoritmu byly výběrové soubory kvantifikovány pro účely tohoto experimentu při v předchozím kroku stanovených jednotkách. Extrahovány byly proměnné na třech jazykových hladinách, které byly ponechány dle původní Menzerathovy a Hřebíčkovy vize. Je však pro další výzkum doporučeno odlišovat hladiny a jednotky při studiu výběrových souborů pro jejich akustické, systematické a grafické vlastnosti. Jedná se tedy o relaci sémantický konstrukt (v délce svých klauzí) – klauze (v průměrné délce svých slov), klauze (v délce svých slov) – slova (v průměrné délce svých slabik a slova (v délce svých slabik) – slabiky (v průměrné délce svých fonémů). V dalších experimentech bude rozšíření počtu zkoumaných hladin vítáno.

V pátém kroku jsou využity statistické a numerické metody k výpočtu parametrů Menzerath-Altmannova zákona, přičemž parametry b_i prokázaných jazykových fraktálů (resp. reciproká hodnota jejich aritmetického průměru) slouží k výpočtu míry sémantičnosti daného výběrového souboru *D*.

V následujícím, šestém kroku je nutné otestovat spolehlivost celého modelu opět pomocí statistických metod. Jedná se o výpočet a posouzení intervalů spolehlivosti a koeficientu determinace.

V sedmém kroku jsem parametry interpretovala ve fraktální analýze a v osmém byly jazykové fraktály pomocí jedné z definic fraktálu vizualizovány společně s přidruženými matematickými fraktály. Jazykové fraktály jsou aproximacemi „dokonalých“ fraktálů matematických.

Na závěr celého algoritmu musí dojít k interpretaci výstupů, tedy k již zmíněnému překladu výstupů exaktních do jazyka lingvistiky. K interpretaci byly připojeny i výsledky shlukové analýzy, které prokázaly, co bylo již předtím heuristicky patrné. Tedy že některé výběrové soubory jsou si z objektivních důvodů „bližší“ než jiné, jmenovitě například překlady originální básně Ottou F. Bablerem do českého a německého jazyka.

Jelikož byla většina výběrových souborů poetických, nabízelo se jejich vyhodnocení pomocí aparátu numerické estetiky a teorie informace. Byly vypočteny entropie, redundance, informační toky, hodnoty překvapení a nápadnosti a estetické míry. Porovnány byly celé výběrové soubory stejně tak jako některé z nich vybrané znaky.

Od doby svého vzniku inspirovala báseň Edgara Allana Poe *The Raven* nejen k překladům a reinterpretacím, ale i k dalším uměleckým zpracováním. Příkladem budiž knižní vazba J. H. Kocmana nebo pozoruhodné zpracování Dalibora Chatrného, viz příloha XI., ve kterém je spojeno několik překladů s originálním textem, čili se v podstatě blíží základní myšlence tohoto experimentu. Na první pohled podobná práce s textem básně předcházela experimentu samotnému, kdy jsem text podle rozličných pravidel zabarvovala a snažila se najít a rozpoznat jeho vnitřní strukturu. Co se zdálo náznakem být patrné, muselo ale být kvantitativně a rigirózně prokázáno. Proto byl vybudován zmíněný algoritmus, proto byly experimentálně otestovány všechny zvolené výběrové soubory. Tento experiment je ale jen prvním, byť důležitým krokem pro celý naznačený budoucí výzkum.

Seznam použité literatury

ACHMANOVOVOVÁ, O. a kol.: *Exaktní metody v jazykovědě*. Praha: SPN 1965.

ALTMANN, G. Prolegomena to Menzerath's Law. *Glottometrika*, 1980, 2, s. 1-10.

ALTMANN, G. – SCHWIBBE, M. H. – KAUMANNS, W. *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Hildesheim: Olms, 1989.

ANDRES, J. On de Saussure's principle of linearity and visualization of language structures. *Glottotheory,* 2009, 2, 2, s. 1-14.

ANDRES, J. On a Conjecture about the fractal structure of language. *Journal of Quantitative Linguistics*, 2010, 17, 2, s. 101-122.

ANDRES, J. – BENEŠOVÁ, M. – KUBÁČEK, L. – VRBKOVÁ, J. Methodological note on the fractal analysis of texts*. Journal of Quantitative Linguistics*, 2011, 18, 4. To appear.

ANDRES, J. – BENEŠOVÁ, M. Fractal analysis of Poe's Raven*. Glottometrics*, 21, 2011, s. 73-100. To appear.

ANDRES, J. – RYPKA, M. Self-similar fractals with a given dimension and the application to quantitative linguistics. *Non-linear Analysis – B (Real World Applications),* 2011. To appear.

BARNSLEY, M. F. *Fractals Everywhere*. New York: Academia Press, 1988.

BARTÓK, I. – JANOUŠEK, I. *Počítače a umenie*. Bratislava : SPN, 1980. 169 s.

BENEŠOVÁ, M. *Artware: Estetické aspekty matematických objektů*. Olomouc, 1999. 93 s. Diplomová práce. Univerzita Palackého Olomoouc.

BENEŠOVÁ, M. *Rudimenty kvantitativní lingvistiky se zvláštním přihlédnutím k teorii o fraktální povaze textu.* Olomouc 2007.

BENEŠOVÁ, M. Numerická estetika, počítače a umění. *Logos Polytechnikos*. 2010, 1, 4, s. 66-83. Dostupný také z WWW:<http://vspj.cz/veda_vyzkum/logos.php?id=4&id_druha_uroven=161>.

BUK, S. – ROVENCHAK, A. Menzerath-Altmann Law for syntactic structures in Ukrainian. *Glottotheory*. 2008, 1, 1, s. 10-17. Dostupný také z WWW: <http://arxiv.org/abs/cs/0701194>.

COVENEY, P. – HIGHFIELD, R. *Mezi chaosem a řádem.* Praha: Mladá fronta 2003.

ČERNÝ, J. *Dějiny lingvistiky*. Olomouc: Votobia, 1996. 517 s.

ČERNÝ, J. *Úvod do studia jazyka*. Olomouc: Rubico, 1998.

DEVLIN, K. *Jazyk matematiky – jak zviditelnit neviditelné*. Praha: Argo a Dokořán 2002.

DUŠKOVÁ, L., et al. *Mluvnice současné angličtiny na pozadí češtiny*. Praha : Academia, 1994.

DVOŘÁKOVÁ, A. Havran vícekrát. Orientace/studovna, *Lidové noviny* (sobota 14.3.2009). Praha: MAFRA, a.s., 21.

EFTEKHARI, A. Fractal geometry of texts: First attempt to Shakespeare's works. *Journal of Quantitative Linguistics*. 2006, 13, 2-3, s. 177-193.

FALCONER, K. *Fractal Geometry: Mathematical Foundations and Applications*. Chichester - New York: Wiley & Sons, Inc., 1990. 155 s.

FALTÝNEK, Dan. *Sémiotické primitivy v gramatické konstrukci*. Olomouc, 2011. 120 s. Dizertační práce. Universita Palackého Olomouc.

FERNAU, H. – STAIGER, L. Iterated function systems. *Information and Computation* 168(2). 2001. s. 125 – 143.

GLEICK, J.: *Chaos. Vznik nové vědy.* Praha: Ando Publishing, 1987.

GUTIÉRREZ, J. M. – COFIÑO, A. S. – ABBOT, P. Challenging the boundaries of symbolic computation. In: *Proceedings of fifth International Mathematical Symposium* (IMS'03, ed. by Mitic, P., Ramsden, P., and Carne, J.). London: Imperial College Press, 2003. s. 1 – 8.

GARCIA, E.: *The Fractal Nature of Semantics.* Dostupné z http://www.miislita.com/factals/factal.html

HAJIČOVÁ, E. – PANEVOVÁ, J. – Sgall, P. *Úvod do teoretické a počítačové lingvistiky*. Praha, Karolinum, 2002.

HEIBEGER, R. M. – HOLLAND, B. Statistical Analysis and Data Display. New York: Springer, 2004.

HŘEBÍČEK, Luděk. Journal of Quantitative Linguistics. *Vesmír* [online]. 1994, 73, 166, [cit. 2011-08-23]. Dostupný z WWW: <http://www.vesmir.cz/clanek/journal-of-quantitative-linguistics>.

HŘEBÍČEK, L. *Text Levels. Language Constructs, Constituents and the Menzerath-Altmann Law*. Trier: Wissenschaftlicher Verlag Trier, 1995.

HŘEBÍČEK, L. *Lectures on Text Theory*. Praha: Oriental Institute, 1997.

HŘEBÍČEK, L. *Variation in Sequences*. Praha: Academia, 2000.

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

HŘEBÍČEK, L. *Text in Semantics: The Principles of Compositness*. Praha: Oriental Institute, 2007.

HŘEBÍČEK, Luděk. Filologie versus lingvistika. *Vesmír* [online]. 2008, 87, 488, [cit. 2011-08-23]. Dostupný z WWW: <http://www.vesmir.cz/clanek/filologie-versus-lingvistika>.

JAŘAB, J. – MASNEROVÁ, E. – NENADÁL, L.: *Antologie americké literatury.* Praha: SPN 1985.

JAIN, A. – DUBES, R. *Algorithms for Clustering Data*. New York: Prentice Hall, Upper Saddle Rivers, 1998.

KÖHLER, R.: Maßeinheiten, Dimensionen und fractale Structuren in der Lingvistik. *Zet-Zeitschrift für Empirische Textforschung* 2, 5-6, 1995.

KUBÁČEK, L. Confidence Limits for Proportions of Linguistic Entities. *Journal of Quantitative Linguistics*. 1994, 1, s. 56-61.

KUBÁČEK, L. – KUBÁČKOVÁ, L. *Statistika a metrologie*. Olomouc: Palacký University Press, 2000.

MANDELBROT, B.: *The Fractal Geometry of Nature*. New York: Freeman, 1982.

MANDELBROT, B.: *Fraktály. Tvar, náhoda a dimenze*. Praha: Mladá fronta, 2003.

NEBESKÝ, P. Investiční životní pojištění je v Česku stále populárnější - vydělává totiž. *Svitanský deník*. 26,10,2009, 26, s. 5.

NOSEK, J. a kol.: *Chaos, věda a filosofie.* Praha: Filosofia, 1999.

ORLOV, J. K. – BORODA, M.G. – NADAREJŠVILI, I.Š. *Sprache, Text, Kunst : Quantitative Analysen*. Bochum: Brockmeyer, 1982.

PAVLÍK, J. Informace, ontologie, entropie. *E-Logos : Electronic Journal for Philosophy* [online]. 2004, č. 4, [cit. 2010-08-09]. Dostupný z WWW: <http://nb.vse.cz/kfil/elogos/epistemology/pavl1-04.pdf>. ISSN 1211-0442.

PEITGEN H. – JÜRGENS, H. – SAUPE, D. *Chaos and Fractals*. New York: Springer, 2004.

PETR, J., et al. *Mluvnice češtiny 1: Fonetika, Fonologie, Morfonologie a morfemika, Tvoření slov*. Praha: Academia, 1986a.

PETR, Jan, et al. *Mluvnice češtiny 2: Tvarosloví*. Praha: Academia, 1986b.

PETR, Jan, et al. *Mluvnice češtiny 3: Skladba*. Praha: Academia, 1987. 748 s. ISBN 21-029-88.

PETRIE, Aviva – WATSON, Paul. *Statistics for Veterinary and Animal Science*. Oxford: Blackwell Publishing, 2006.

POE, E. A. *Der Rabe.* Übersetzt und herausgeben von Otto F. Babler. Olmütz: Heiliger Berg bei Olmütz, 1931.

POE, E. A. *Havran : Šestnáct českých překladů*. Praha: Odeon, 1985.

POE, E. A. *Havran krkavec*. Praha: Lyra Pragensis, 1993.

POE, E. A.: *Spirit of the Dead: Tales and Poems.* London: Penguin Popular Classics, 1997.

POE, E. A. *Havran. Devet prekladov do slovenčiny.* Bratislava: Petrus, 2004.

POE, E. A. *Krkavec/The Raven*. Praha: Aleš Prstek, 2008a.

POE, E. A. The Raven*. Literární revue Weles*, 2008b. s. 32-33.

RALSTON, A. *A First Course in Numerical Analysis*. New York: McGraw-Hill, 1965.

SAUSSURE, F. de. *Kurz obecné lingvistiky*. Praha: Academia, 2007.

SGALL, P. – BÉMOVÁ, A. – BENEŠOVÁ, E. – GORALČÍKOVÁ, A. – HAJIČOVÁ, E. – MACHOVÁ, S. – PANEVOVÁ, J. – PIŤHA, P. – ŘÍHA, A. – VOMÁČKA, I. – WEISHEITELOVÁ, J. *Úvod do algebraické lingvistiky.* Praha: SPN, 1973.

SHANNON, B. Fractal patterns in language. *New Ideas in Psychology* 11(1). 2009. s. 105 – 109.

SVATOŠOVÁ, L. – KÁBA, B. *Statistické metody I*. Praha: ČZU Praha, 2009.

STOER, J. – BULIRSCH, R. *Introduction to Numerical Analysis*. New York: Springer, 2002.

STRUIK, D. J. *Dějiny matematiky*. Praha: Orbis, 1963.

ŠTEKAUER, P. et al. *Rudiments of English Linguistics*. Prešov: Slovacontact, 2000.

TĚŠITELOVÁ, M. *Kvantitativní lingvistika*. Praha: SPN, 1987a.

TĚŠITELOVÁ, M.: *O češtině v číslech*. Praha: Academia, 1987b.

TĚŠITELOVÁ, M.: *Quantitative Linguistics*. Praha: Academia, 1992.

VOLÍN, J. *Statistické metody ve fonetickém výzkumu*. Praha: Epocha, 2007.

WILDGEN, W. Chaos, fractals and dissipative structures in language. In: ALtmann, G. & Koch, W. A. (eds.). *Systems. New Paradigms for the Human Sciences*. Berlin: de Gruyter, 2011. s. 596 – 620.

WIMMER, G, et al. *Úvod do analýzy textov*. Bratislava: Veda, 2003.

[1] *Kybernetika a umělá inteligence* [online]. 2000 [cit. 2010-08-08]. Teorie informace a entropie. Dostupné z WWW: <cyber.felk.cvut.cz/gerstner/teaching/kui/sbirka/1_TeorieI.doc>.

[2] Perry, Lynellen D.S. *Research Topic Approval*. Dostupné z http://www.lynellen.com/write/restopic.html .

[3] Pollard-Gott, L. *Fractals in Poetry*. Dostupné z http://classes.yale.edu/fractals/IMA/FB/ArtFrac/FractalPoetry.html

[4] Leopold, E. *Fractal Structures in Language. The Question of the Imbedding Space*. Dostupné z http://www.mt.haw-hamburg.de/home/leopold/publist/hrebinet.ps

[5] *Dalibor Chatrný* [online]. 2011 [cit. 2011-08-29]. E. A. Poe: Havran. Dostupné z WWW: <http://www.chatrny.cz/v/PraCeNaPapiReTextilii1990-1999/PoeziePrekladyAtd/Havran1996_420X590_BarevneTuze/1996_420X590_BarevneTuzeTuzkaPapiR_1.jpg.html>.

Seznam příloh

Přílohy