Filozofická fakulta Univerzity Palackého v Olomouci
Katedra obecné lingvistiky

Výňatek z historického vývoje užívání Zipfových zákonů v lingvistickém kontextu

magisterská diplomová práce

Autor: Bc. Eliška Syrovátková
Vedoucí práce: Mgr. Lukáš Zámečník, Ph.D.

Olomouc
2016

Prohlášení

Prohlašuji, že jsem magisterskou diplomovou práci „Výňatek z historického vývoje užívání Zipfových zákonů v lingvistickém kontextu“ vypracovala samostatně a uvedla jsem veškerou použitou literaturu a veškeré použité zdroje.

V Olomouci dne 22. 8. 2016 Podpis

Poděkování:

Děkuji doktoru Zámečníkovi, který je neocenitelným odborným rádcem, vždy ochotným vedoucím a trpělivým člověkem. Velmi mu děkuji za vedení této práce, jeho připomínky a rady. Za úvodní korekturu děkuji Adéle Hazuchové, za obětavou závěrečnou spolupráci Báře Hrabalové. Ale především děkuji výjimečné Evě Syrovátkové, bez které by nic z tohoto nebylo možné.

Abstrakt

Název práce: Výňatek z historického vývoje užívání Zipfových zákonů v lingvistickém kontextu

Autor práce: Bc. Eliška Syrovátková

Vedoucí práce: Mgr. Lukáš Zámečník, Ph.D.

Počet stran a znaků: 121 170

Počet příloh: 0

Abstrakt: Tématem práce jsou Zipfovy zákony a vývoj jejich vnímání. Práce se zaměřuje na analýzu pomocí Zipfových zákonů na písmenech, slovech, frázích, větách, textech, korpusech, komunikaci a náhodně vygenerovaných textech. Představuje odlišné vědecké závěry na tato jednotlivá lingvistická témata, sama se nesnaží na položené otázky najít konečnou odpoveď. Vychází z textů George Kingsleyho Zipfa, Luďka Hřebíčka, Wentiana Li, Jakea Rylanda Williamse nebo Ramona Ferrer-i-Cancha a Brity Elvevåg. Práce obsahuje i životopis George Kingsleyho Zipfa a popis vzniku jeho zákonů. Také se dotkne vymezení pojmu slova a určení jeho délky pro potřeby matematicko-lingvistické analýzy textu. V práci jsou krátce představeny i metody generování náhodného textu v českém jazyce spolu s popisem veřejně dostupných programů na generování náhodných textů v českém jazyce. Text je kompilací článků na téma Zipfových zákonů, které byly publikovány v anglickém jazyce.
Klíčová slova: Zipfův zákon, náhodné texty, fráze, matematická lingvistika

Abstract

Title: An excerpt from the historical development of the usage of Zipf´s law in linguistic context

Author: Bc. Eliška Syrovátková

Supervisor:Mgr. Lukáš Zámečník, Ph.D.

Number of pages and characters: 121 170

Number of appendices: 0

Abstract: The topic of the thesis is the Zipf's law and its usage. The thesis focuses on the use of the Zipf's laws in analyzing letters, words, phrases, sentences, texts, corpuses, communication and randomly generated texts. It presents various scientific conclusions for particular linguistic topics, it does not attempt, however, to find a final answer to the set questions. It is based on the texts of George Kingsley Zipf, Luděk Hřebíček, Wentian Li, Jake Ryland Williams or Ramon Ferrer-i-Canch a Brita Elvevåg. The thesis also contains George Kingsley Zipf's biography and the description of his laws. It also touches on the definition of the term word and estimation of its length for the purposes of mathematical-linguistic text analysis. The methods of generating random texts in the Czech language are also introduced in the thesis, along with the description of the publicly accessible random text generators. The text is a compilation of articles focused on the topic of the Zipf's laws, that were published in English.

Keywords: Zipf's law, random texts, phrases, mathematical linguistic

Obsah

Úvod

„V průběhu minulého století bylo zjištěno, že prvky mnoha různých systémů přibližně následují Zipfův zákon distribuce - od městské populace, přes velikost firem až po příjmení."[1] Tento namátkový výběr příkladů ilustruje, jak rozmanité jsou oblasti lidského života, které už někdy byly posuzovány optikou jednoho původně matematicko-lingvistického zákona.

George Kingley Zipf byl lingvista, který se studium jazyka snažil přiblížit přírodním vědám. Ve 30. letech minulého století zformuloval trojici matematizovaných tezí o lidském jazyce, pro které se vžilo pojmenování *Zipfovy zákony*. Konkrétně se sice týkaly jazyka, ale řešily (a trochu zjednodušovaly) jeho strukturu jakožto nesmírně složitého systému. Proto bylo zřejmě lákavé si tyto principy vypůjčit a pokusit se je ověřit i na jiných nesmírně složitých systémech. Za těch skoro 90 let, které uplynuly od prvního vydání Zipfovy práce, se s těmito principy pracovalo v opravdu široké škále oborů, byly potvrzovány i vyvraceny.

Tato práce se vrací k jejich původnímu zaměření, tedy k jazykovému. V jazykovědě se tomuto tématu během let věnovalo velké množství vědců, kteří se zaměřovali na různé oblasti jazyka a docházeli k různým výsledkům a závěrům. Není v silách diplomové práce obsáhnout vývoj okolo Zipfových zákonů v celé jeho komplexnosti a tato si to ostatně za cíl ani neklade. Dílčím cílem diplomové práce je představení autora Zipfových zákonů a stručné načrtnutí osobností, které na jeho práci navazovaly. Hlavním cílem této práce je představit některé směry bádání na základních jazykových jednotkách jakými jsou písmena, slova, ustálená slovní spojení, texty, korpusy textů, samotná komunikace a závěr práce je věnován specifickému lingvistickému tématu - náhodně generovaným textům. Ty ze své podstaty narušují představu Zipfových zákonů coby vhledu do struktury složitého systému přirozeného lidského jazyka. Tedy za předpokladu, že vykazují známky Zipfových zákonů, na čemž se vědci neshodnou. I když jsou použité zdroje z valné většiny v anglickém jazyce a tedy reflektují spíše výsledky na angličtině, téma náhodně generovaných textů je v českém prostředí o něco specifičtější, proto se u něj práce na jednu podkapitolu zastaví.

[1] WILLIAMS, Jake Ryland, LESSARD, Paul, DESU, Suma, CLARK, Eric, BAGROW, James, DANFORTH, Christopher, DODDS, Peter Sheridan. Zipf's law holds for phrases, not words. In *Scientific reports 5*, 2015, str. 1.

Celá práce je koncipována jako komparativní kompilace vědeckých textů z anglofonního prostředí, která dává pouze nahlédnout na vývoj myšlenek a na protichůdné závěry vědeckých studií. Neposkytne konečnou odpověď a v žádném případě nepředkládá téma v celé jeho komplexnosti, pouze z něj vybírá a sestavuje avizovaný výňatek z dlouhé cesty, kterou Zipfovy zákony od doby svého vzniku ušly.

1 George Kingsley Zipf a jeho zákony

Kvantitativní lingvistika bývá někdy označována za Zipfovu lingvistiku, protože právě George Kingsley Zipf je považován za jejího zakladatele. Pokusil se postavit jazykovědu na roveň přírodním vědám zavedením kvantitativních metod zkoumání. I když podle Ronalda Wyllyse byla Zipfova znalost matematiky minimální a povědomí o statistice zřejmě neexistující, neboť údajně mnohem více energie vkládal do filozofování o implikacích svých principů.[2] Za tímto postřehem samozřejmě nestojí snaha zdiskreditovat Zipfa samotného či jeho práci a odkaz.

George Kingsley Zipf[3]

[2] WYLLYS, Ronald. Empirical and theoretical bases of Zipf's law. In *Library Trends 30 (1) Summer 1981*, str. 57.
[3] Fotografie převzata z *Glottometrics 3, 2002*

1.1 Výňatky ze Zipfova života

George Kingsley Zipf se narodil 7. ledna 1902 ve státě Illinois, konkrétně ve městě Freeport. Z otcovy strany má sice německé kořeny, jeho dědeček Frederic Sebastian Zipf přišel z Německa do Ameriky roku 1850, ale podle Zipfova nejstaršího syna jej to nijak nepředurčovalo k jeho budoucímu zájmu o německý jazyk – Frederic Zipf zastával pevné stanovisko, že z rodiny se stali Američané, proto budou mluvit pouze anglicky.[4]

George K. Zipf byl vždy dobrým studentem – už na střední škole Freeport High School vyhrál cenu náležející žákovi s excelentními studijními výsledky a dál ve svém vzdělávání pokračoval na Harvard College,[5] kterou roku 1924 ukončil rovněž s nejlepšími výsledky.[6] Po promoci odjel studovat na berlínskou univerzitu, kde jej poprvé napadlo zkoumat jazyk jako přírodní jev, v čemž pokračoval i po návratu do Států. Posléze získal na Harvardu doktorát z komparativní filologie.[7] „Při studiu fonetických změn v jazycích se začal zajímat o frekvenci užití fonémů jakožto faktoru, který ovlivňuje jejich sklon k fonetickým změnám během delších časových období.[8] Od relativních frekvencí fonému se přesunul ke studiu relativních frekvencí slov a v roce 1932 vydal knihu *Selected studies of the principle of relative frequency in language*.“[9] V této knize použil Lotkův zákon, ale řádně na něj neodkázal, a proto je tento některými lidmi považován za jeden ze Zipfových zákonů.[10]

---

[4] PRÜN, Claudia, ZIPF, Robert. Biographical notes on G. K. Zipf. In *Glottometrics 3, 2002*, str. 3.

[5] Jedna ze škol v rámci Harvardovy univerzity, na níž je možno získat bakalářský titul.

[6] PRÜN, Claudia, ZIPF, Robert. Biographical notes on G. K. Zipf. In *Glottometrics 3, 2002*, str. 3.

[7] Tamtéž.

[8] Na toto téma ostatně v roce 1929 napsal svou disertační práci Relative frequency as a determinant of phonetic change.

[9] WYLLYS, Ronald. Empirical and theoretical bases of Zipf's law. In *Library Trends 30 (1) Summer 1981,* str. 56.

[10] ROUSSEAU, Ronald. George Kingsley Zipf: life, ideas, his law and informetrics. In *Glottometrics 3, 2002,* str. 13.

Zipf začal na Harvardově univerzitě vyučovat německý jazyk a nadále pokračoval ve zkoumání jazyků.[11] Jeho práce byla veskrze interdisciplinární a on se tak ve svém akademickém působení čím dál více přikláněl i ke společenským vědám.[12] V roce 1935 publikoval další knihu *The psycho-biology of language*, kde se objevil i graf vztahu frekvence-rank pro slova v latinských textech Tita Maccia Plauta. Svůj zájem o společenskovědní témata a společenské fenomény pak promítl do knihy *National unity and disunity*.

Ve třicátých letech se George Kingsley Zipf oženil s Joyce Waters Brownovou a postupně v onom desetiletí měli čtyři děti – Roberta, Katherin Slater, Joyce Bogardus a Henryho. Nejprve žili v Cambridgi, kde se ostatně manželé potkali, později se rodina přestěhovala do Newtonu, kde děti začaly chodit do školy.

„Během druhé světové války byl požádán, aby se přestěhoval do Washingtonu a svou prací vypomohl válečnému úsilí, ale on odmítl a zůstal na Harvardu. V té době vedl svůj první ročník němčiny na Harvardu a obracel důraz na učení slovní zásoby, aby studenti byli na konci jednoho roku vysokoškolské němčiny schopni číst německy bez pomoci slovníku. (…) Vycházel z předpokladu, že studenti pravděpodobně nebudou mít několik let na to, aby je věnovali studiu jazyka na Harvardu, místo toho se po současném školním roce přidají k armádním silám, nebo je odvodová komise povolá do armády, až školní rok skončí.“[13]

Svou poslední knihu *Human behavior and the principle of least effort: An introduction to human ecology* vydal Zipf v roce 1949. V následujícím roce obdržel Guggenheimovo stipendium pro kvantitativní studium určitých tržních fenoménů za účelem odhalení skrytých statistických pravidelností.[14] „Zipf plánoval využít svobodu zaručovanou Guggenheimovým stipendiem k tomu, aby se věnoval svému plánu prozkoumat americké podnikání, jeho firmy a jejich spravování. (…) Tento směr se projevil i v několika jeho pracích publikovaných v letech 1949 a 1950,

[11] Aby mohl Zipf analyzovat texty v různých jazycích, potřeboval pomoc svých studentů, kteří ho měli poměrně v oblibě, avšak zpočátku jeho teze nebrali příliš vážně.

[12] PRÜN, Claudia, ZIPF, Robert. Biographical notes on G. K. Zipf. In *Glottometrics 3, 2002*, str. 4.

[13] PRÜN, Claudia, ZIPF, Robert. Biographical notes on G. K. Zipf. In *Glottometrics 3, 2002*, str. 3.

[14] Tamtéž, str. 4.

které analyzují rozličné aspekty amerického podnikání."[15] Guggenheimovo stipendium bylo podmíněno lékařskou prohlídkou a právě při ní lékaři poprvé zjistili, že Zipf má rakovinu, tou dobou už ve velmi pokročilém stadiu. V červnu 1950 sice podstoupil operaci, ale už se pro něj nedalo nic udělat. Z nemocnice se vrátil domů, kde 25. září 1950 zemřel.[16]

1.2 Zipfovy zákony

Pokud se hovoří o Zipfových zákonech, jsou tím míněna tři pravidla, která představil ve své knize z roku 1949 *Human behavior and the principle of least effort* v kapitole *On the economy of words*. Někdy se však používá pouze pojem Zipfův zákon, čímž se z oněch tří míní ten první a nejznámější.

První zákon stanovuje vztah frekvence slov a jejich ranku, tedy, že jejich součin je konstantou -> $r \cdot f = C$. Rank je pořadí slova ve frekvenčním slovníku – čím nižší rank, tím vyšší frenkvence, přičemž jednomu ranku náleží pouze jedna frekvence. Zipf vycházel z analýzy knihy Jamese Joyce *Ulysses*, pro který vytvořil frekvenční slovník zhruba třiceti tisíc slov. „I když nelze snadno v tabulce zobrazit vztah ranku a frekvence pro všechna tato rozmanitá slova, přesto je můžeme celkem pohodlně zobrazit v grafu, protože víme, že tato rovnice, tedy $r \cdot f = C$, se v grafu v logaritmickém měřítku zobrazí jako sled bodů svažujících se v přímé linii zleva doprava v úhlu 45°."[17]

Druhý Zipfův zákon se zaměřuje na vztah mezi množstvím slov o jisté frekvenci a právě oné jejich frekvenci. "Je tu poměr nepřímý, neboť stoupá-li koeficient frekvence, klesá počet slov, která tuto frekvenci mají."[18] Tento nepřímý poměr je pak vyjádřen formulí, kde počet slov o dané frekvenci krát frekvence na druhou je konstantní -> $a \cdot b^2 = k$. Zipf však tomuto svému zákonu nepřikládal obecnou platnost, omezil ji pouze na slova s nevelkou frekvencí, která představují podstatnou

[15] PRÜN, Claudia, ZIPF, Robert. Biographical notes on G. K. Zipf. In *Glottometrics 3, 2002*, str. 4.
[16] Tamtéž.
[17] ZIPF, George. *Human behavior and the principle of least effort.* 1949, str. 24.
[18] TĚŠITELOVÁ, Marie. K statistickému výzkumu slovní zásoby. In *Slovo a slovesnost*, r. 22, č. 3, 1961.

část slovníku. "Podle Zipfa lze tímto zákonem zjistit značný stupeň pravidelnosti v distribuci slov, která svědčí o tendenci v jazyce udržet rovnováhu mezi frekvencí slov a počtem slov různých."[19] Na českém materiálu Marie Těšitelová ověřovala[20] platnost tohoto zákona a ukázalo se, že "tento zákon nejen neplatí o slovech s vysokou frekvencí, nýbrž i o slovech s frekvencí poměrně nízkou."[21]

Třetím zákonem Zipf popisuje vztah mezi frekvencí slova a počtem jeho významů -> $\frac{m}{\sqrt{f}} = k$. Ve vzorci m odpovídá počtu významů slova, čímž z něj v praktickém užití vyplývá, že slova s vyšší frekvencí by měla mít více významů.

2 Přímé reakce na Zipfovy zákony

Do kapitoly přímých reakcí jsou zařazeni autoři navazující na Zipfovu práci a články publikované do roku 1960, které jejich tvůrci nezamýšleli pouze jako aplikaci Zipfových tezí na různé části jazyka, ale především jako přímou polemiku s tvrzeními, která George Kingsley Zipf ve svých pracích představil. Někteří tyto teze odmítli, jiní je naopak rozšířili či opatřili dalšími komentáři.

2.1 Edward Lee Thorndike

Tento americký psycholog a pedagog, žijící v letech 1874–1949, byl průkopníkem behaviorismu a etologie. Vystudoval Harvard, poté získal na Kolumbijské univerzitě doktorát a zůstal na ní jako pedagog. Ve svých experimentech se zabýval především procesem učení,[22] pojmenoval zákon přirozeného učení a zavedl

[19] TĚŠITELOVÁ, Marie. K statistickému výzkumu slovní zásoby. In *Slovo a slovesnost*, r. 22, č. 3, 1961.
[20] TĚŠITELOVÁ, Marie. *Otázky frekvence slov (zvláště v češtině)*, nepublikovaná disertační práce, 1951, 10-22.
[21] TĚŠITELOVÁ, Marie. K statistickému výzkumu slovní zásoby. In *Slovo a slovesnost*, r. 22, č. 3, 1961.
[22] Thorndike vytvářel pokusy s puzzle boxy, do kterých umisťoval zvířata, která měla za úkol přijít na fungování mechanismu, s jehož pomocí se dostala buď na svobodu nebo za potravou.

pojem instrumentální podmiňování, tedy učení pokusem a omylem. Z tohoto svého zaměření se chvíli věnoval i vytváření knih[23] vhodných k učení čtení psaní, které sestavoval použitím co nejběžnějších a nejfrekventovanějších slov ve standardní angličtině.

2.2 Benoît Mandelbrot

Francouzsko-americký matematik, který se roku 1924 narodil v litevsko-židovské rodině ve Varšavě, studoval v Paříži na École Polytechnique, poté dva roky přes stipendium na kalifornském Caltechu, a nakonec jako postgraduální asistent pracoval na univerzitě v Princetonu. Mandelbrot byl především matematik a zakladatel oboru fraktální teorie, věnoval se však i ekonomii, teorii informace, podílel se na vzniku prvních počítačových programů, a zabýval se dokonce vývojem cen na finančních trzích, termodynamikou či turbulencí kapalin.

Mandelbrot strávil část své kariéry v IBM, kde nastoupil jako konzultant pro kolegy zaměřené na praktické problémy, díky čemuž nahlédnul a zasahoval do více oborů. Mimo jiné jej to přivedlo také k šumu a chybovosti na telefonních linkách. A právě jeho práce z roku 1953 *An informational theory of the statistical structure of language* se primárně zabývá přenosem informace kanálem, u kterého je třeba zohlednit strukturu jazyka a tím i zprávy. Také se ve svém textu soustředí na způsob, jakým se v mozku informace kóduje a úsporná kritéria toho, aby zmíněná struktura i způsob kódování sobě odpovídaly.[24] V článku proto pracuje s Shannonovými teoriemi, a zároveň zasazuje do kontextu moderní teorie komunikace práci Saussurovu. K Zipfovu pojetí jazyka se dostane v části, kdy výsledky předchozích diskuzí označí za nezávislé na jazyku – „proto kvantitativní výsledky statistických struktur konkrétních entit, které mají být použity jako kritérium pro jejich identifikaci mezi nejrůznějšími

[23] Například *A teacher's word book of the twenty thousand words found most frequently and widely in general reading for children and young people* z roku 1932 nebo *The teacher's word book of 30,000 words* z roku 1944

[24] MANDELBROT, Benoît. An informational theory of the statistical structure of language. In *JACKSON, Willis. Communication theory: Papers read at a symposium on "Applications of communication theory" held at the Institution of electrical engineers, London September 22nd-26th 1952.* 1953, str. 486–487.

jednotkami informace, musí být stejné pro všechny jazyky. Matematické podrobnosti vlastně ukazují, že všechny varianty kritéria nejmenšího úsilí vedou ke stejné ‚kanonické' rodině zákonů pro konkrétní entity."[25] Mandelbrot proto začal pracovat se vztahem ranku a frekvence slov, který je podle něj vždy ovlivněn tím, jak si výzkumník na začátku definuje pojem slova. V případech, kdy je slovo jednoznačně definováno, je pak možno teorii vztahu ranku a frekvence použít jako diskriminační pravidlo pro výběr té lepší možnosti z mnoha způsobů, jak rozdělit řadu fonémů do slov.[26]

„Ale Zipfův výklad nebyl ani v nejmenším založen na jakékoliv seriózní lingvistické teorii ani na žádné komunikační teorii."[27] Proto Mandelbrot upravuje formuli vztahu ranku a frekvence ve tvaru $P_n = P_n - 1$, kterou „představil J. B. Estoup a dále promyslel Zipf"[28] přidáním „dvou parametrů, *m* a *B*, které představují nezbytné vylepšení, protože počítají s poměrně výraznými nesrovnalostmi s hlavním trendem reprezentovaným Zipfovým zákonem."[29] Upravenou formuli ve tvaru $P_n = P_{(n+m)} - B$ porovnává v grafu s původní formulí:

[25] MANDELBROT, Benoît. An informational theory of the statistical structure of language. In *JACKSON, Willis. Communication theory: Papers read at a symposium on "Applications of communication theory" held at the Institution of electrical engineers, London September 22nd-26th 1952.* 1953, str. 490–491.

[26] Tamtéž, str. 491.

[27] Tamtéž, str. 492.

[28] Tamtéž.

[29] Tamtéž.

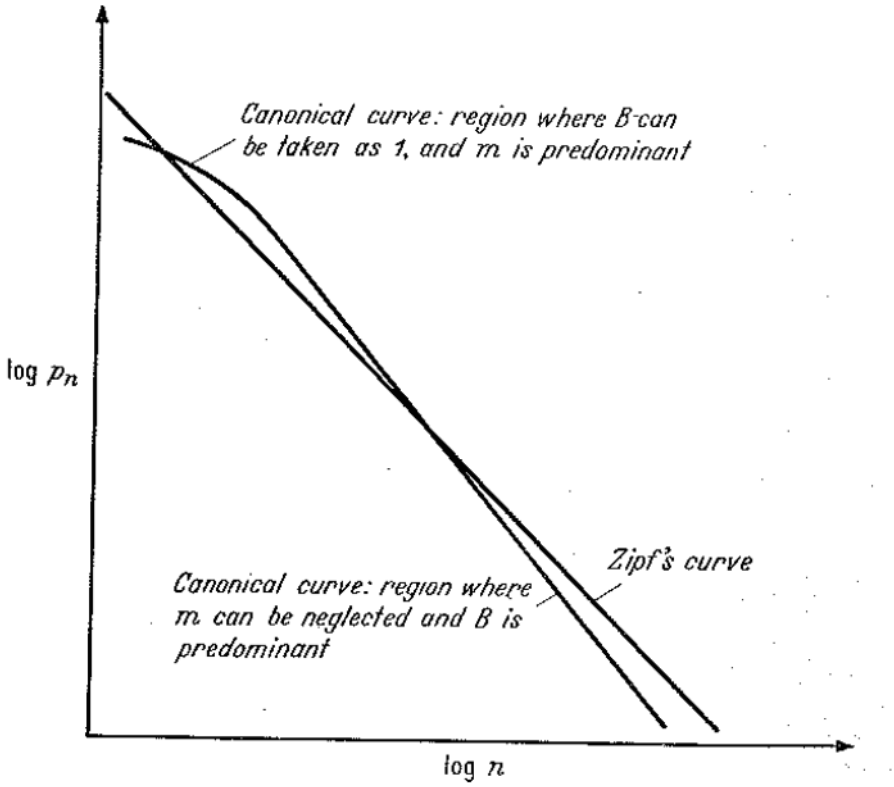


Figure 2. Canonical Rank-frequency distribution $p_n = P(n + m)^{-B}$ as compared with Zipf's "law" $pn = P^{n-1}$

Hlavním záměrem článku a výpočtů v něm obsažených je, jak uvedl sám Mandelbrot, ukázat, že jazyk je kanonická zpráva. Autor text psal primárně pro oči komunikačních techniků, ne filologů.[30] „Neaplikuji kanonický zákon na slova, ale zjistil jsem, že Zipfova data s plně flektivními slovy jsou věrně reprezentována kanonickou statistikou."[31]

Mandelbrot se k Zipfovi opět vrátil o rok později v práci *Simple games of strategy occuring in communication through natural languages*, jejímž hlavním tématem byl opět přenos jazykové informace, tentokrát obohaceno o řešení situací, kdy je nutno posílanou informaci zašifrovat a následně rozšifrovat. V tomto článku

[30] MANDELBROT, Benoît. An informational theory of the statistical structure of language. *In JACKSON, Willis. Communication theory: Papers read at a symposium on "Applications of communication theory" held at the Institution of electrical engineers, London September 22nd-26th 1952.* 1953, str. 502.

[31] Tamtéž.

o Zipfově formuli ve tvaru $p_r = \frac{P}{r}$ tvrdí, že ačkoli se snaží reprezentovat všechny případy, je validní pouze v mezních stavech.[32]

2.3 Herbert Alexander Simon

Američan, který se zajímal o ekonomii, kognitivní psychologii a filozofii žil v letech 1916– 2001. V padesátých letech se svými kolegy představil kritiku teorie objektivně racionálního rozhodování, jehož podle nich nelze dosáhnout, protože v takové situaci jsou kladeny nesplnitelné nároky na kognitivní schopnosti člověka, který se má rozhodnout. V roce 1978 obdržel Nobelovu cenu v oboru ekonomie za „průkopnický výzkum o rozhodovacím procesu v ekonomickém uspořádání."[33]

Ve své práci *On a class of skew distribution functions* se zabývá analýzou distribucí funkcí, které se objevují v širokých škálách empirických dat v sociologii, biologii a výzkumech ekonomických fenoménů. „Objevují se tak často a v tak rozdílných jevech, že to jednoho nutí k domněnce, že pokud mají tyto jevy nějaké shodné charakteristiky, může to být způsobeno pouze podobností ve struktuře skrytých pravděpodobnostních mechanismů."[34] V části věnované distribuci slov vychází Simon z předpokladu, že slovní zásoba jedince je výsledkem stochastického procesu. Vytváří teoretický příklad knihy, která se právě píše a dosáhla délky *k* slov. Stanovíme *f(i, k)* počet rozdílných slov, která se objevují přesně *i* krát v prvních *k* slovech. Pokud je v knize 407 různých slov, která se objeví právě jednou, pak *f(1,k)* = 407. „Frekvence poroste proporcionálně s *k.* Pokud je konkrétní frekvence ‚příliš velká' v porovnání s další nižší frekvencí, poroste pomalejším tempem než je průměr; pokud je ‚příliš malá', poroste rychleji než je průměr."[35]

[32] MANDELBROT, Benoît. Simple games of strategy occuring in communication through natural languages. 1954, str. 134.
[33] The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 1978". Nobelprize.org. Nobel Media AB 2014. Web. 20 Apr 2016. <http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/1978/>
[34] SIMON, Herbert. On a class of skew distribution functions. 1955, str. 425.
[35] Tamtéž, str. 429.

2.4 George Armitage Miller

Americký psycholog, žijící v letech 1920–2012, byl předním představitelem kognitivní psychologie, zabýval se i psycholingvistikou a obecně kognitivními vědami. Nejznámější je svou prací o krátkodobé paměti *The magical number seven, plus or minus two: some limits on our capacity for processing information*. Zákonitost, že v krátkodobé paměti udržíme průměrně 7 (+ – 2) informací, bývá někdy uváděna jako Millerův zákon.

3 Zipfovy zákony ve větách, textech, komunikaci

George Kingsley Zipf ve své původní práci vycházel z toho, jakým způsobem lidé pracují se slovy v širší souvislosti celkové komunikace v rámci jazyka. Proto se tato kapitola věnuje článkům, jejichž autoři se při práci se Zipfovými zákony zaměřují na větší jazykové jednotky, jako jsou věty, celé texty, popřípadě specifické komunikační situace. Analýzám na rozsáhlejších jazykových jednotkách se tato kapitola věnuje v obecné rovině. Neklade důraz na rozbory úzce zaměřených textů, jak jsou kupříkladu konkrétní knihy či prozkoumávání textů, které vytvořil jediný autor, nebo dokonce porovnávání děl různých autorů mezi sebou. Budou představeny články, jejichž záměrem je buď demonstrovat možné nástrahy při snaze potvrdit nebo vyvrátit Zipfovo rozložení, zjistit, zda není v samotné podstatě textů zakotveno, že musejí odpovídat principům formulovaných Georgem Kingley Zipfem, nebo jak vypadají texty, které tyto principy dodržují zcela důsledně.

3.1 Analýza textu pomocí Zipfových zákonů

V roce 2002 vyšlo třetí číslo časopisu *Glottometrics* orientovaného na kvantitativní výzkum jazyka a textu, a to ve formě pocty Georgeovi Kingsleymu Zipfovi u příležitosti stého výročí jeho narození. Mimo jiné zde byl poprvé uveřejněn kompletní soupis Zipfovy bibliografie čítající šest monografií a třicet šest článků. V časopise vyšel i článek českého lingvisty Luďka Hřebíčka s názvem *Zipf's law and text*. A právě ten otevře kapitolu o Zipfových zákonech na větších jazykových jednotkách, protože Luděk Hřebíček ve svém textu ideálně shrnuje, nač je třeba pamatovat při takovém badatelském postupu.

„Zipfův zákon může být chápán jako obecný popis lexikální struktury náležející přirozeným jazykům. Sám Zipf předložil velmi důmyslné vysvětlení onoho zákona v mnohem obecněji pojaté práci *Principle of least effort*, čímž může být považován za zákon vystihující lidské chování." [36] Podle Hřebíčka to znamená, že mentální operace s významy, jak s nimi pracuje lingvistika, jsou zredukovány na vztah ranku a sestupné seřazení frekvencí lexikálních jednotek v jazykovém korpusu. [37] Z lingvistického hlediska však na tom Hřebíčka zaráží podstatná věc: „(…) propast mezi lexikální jednotkou a korpusem naznačuje stěží přípustnou absenci jevu relevantního pro lexikální strukturu, která se skutečně v textech vyskytuje. Jakákoli lexikální jednotka přirozeného jazyka je intuitivní abstrakce, jejíž význam je odvozen od jejího užití v textu. Skutečná sémantická vlastnost slova může být zaznamenána přiměřeněji, pokud jsou slova testována jako jednotky samostatných textů."[38]

Proto je třeba si podle Hřebíčka položit dvě zásadní otázky – je přístup opírající se o korpusy schopen vyjádřit skutečnou lexikální strukturu jazyka? A když se vezmou v potaz struktury samostatných textů, jsou teoretická tvrzení spojená se Zipfovými zákony natolik silná, aby ustála šok způsobený přesunem těžiště testování z korpusu na texty samotné?[39]

---

[36] HŘEBÍČEK, Luděk. Zipf's law and text. In *Glottometrics 3, 2002*, str. 27.

[37] HŘEBÍČEK, Luděk. Zipf's law and text. In *Glottometrics 3, 2002*, str. 27.

[38] Tamtéž.

[39] Tamtéž.

Aby Luděk Hřebíček získal odpověď, zaměřil se nejprve na strukturu textu. Především je podle něj zcela pochopitelná skepse rozprostírající se kolem práce s tak vágním termínem, jakým struktura textu bezpochyby je. Především se jí však vždy míní vzájemné vztahy mezi jazykovými jednotkami, jejichž dvě základní charakteristiky Hřebíček zdůrazňuje. Pokud tedy ve struktuře textu hovoříme o jazykových jednotkách, které se mezi sebou dostávají do vztahu, pak jde buď o „jakoukoli sledovanou nepřerušovanou zvukovou sekvenci jazykových funkcí jakožto nositele přerušované sekvence symbolů kódu na různých úrovních jazyka", nebo mohou být tyto jednotky identifikovány tak, že na různých úrovních jazyka „mohou být popsány jako množina vyznačující se vzájemnou podobností".[40]

Záměrem těchto úvah je „vyobrazení textu jakožto škálované jazykové jednotky, ve které je jazykový soubor (nebo kód) vykreslen do komplikované mřížky získané pomocí rozškálování (nebo segmentace) nepřerušovaného jazykového procesu do jednotek a jejich úrovní. To se týká také jazykových jednotek a jejich významů, které jsou spojeny s jejich umístěním v mřížce."[41] Obojí je podle Hřebíčka důsledkem platnosti na mnoha jazycích ověřeného Menzerath-Altmannova zákona založeného na pojmech jazykového konstruktu a jeho konstituentů,[42] jehož slovní formulace zní „čím delší je jazykový konstrukt, tím kratší jsou jeho komponenty (konstituenty)".[43]

Hřebíček se dále v oddíle struktury textu věnuje nadvětným celkům, které nazývá širším kontextem a činí tak spíše pro potřeby daného textu, neboť v poznámce se dočteme, že se jedná o *hreb*, kteréžto označení pro tento jev používá Hřebíček běžně ve svých ostatních pracích. A každá tato nadvětná struktura je taktéž tvořena konstituenty, jimiž jsou v tomto případě věty určitého textu, v nichž se vyskytují předem dané lexikální jednotky. Zúženým kontextem je pak věta tuto jednotku obsahující. To vše je v této kapitole zmíněno proto, že s nadvětnými strukturami souvisejí podobná omezení při analýze textu, jaká čekají na badatele při snaze rozškálovat text do výše zmíněné mřížky za použití jakýchkoli jazykových

[40] HŘEBÍČEK, Luděk. Zipf's law and text. In *Glottometrics 3, 2002*, str. 28.

[41] Tamtéž.

[42] Konstrukt je definován jako jazyková jednotka na vyšší jazykové úrovni a konstituent jako jednotka na nižší jazykové úrovni. Konstrukt proto lze dělit na konstituenty.

[43] HŘEBÍČEK, Luděk. Zipf's law and text. In *Glottometrics 3, 2002*, str. 28.

jednotek a současně za předpokladu, že platí definování jednotky jejím vztahem k ostatním jednotkám textu a tedy jejím funkcím v kontextu.

„Ten důvod tkví v použitém nástroji k pozorování, tedy ve statistice. Extrémně krátké texty (sestávající se kupříkladu z jednoho či několika slov, z jedné věty nebo z malého množství vět) nejsou vhodné pro taková pozorování. Z pochopitelných důvodů však ani mimořádně dlouhé texty nemohou být za podobným účelem využity; v nepřiměřeně velkých textech mizí detaily lexikálních jednotek vyplývající ze souvislostí a nemohou se tak odrazit ve statistických datech. Pokud text přerůstá nějaké rozumné meze, rozložení lexikálních jednotek podstatně mění svůj směr."[44] Totéž platí v případech, kdy jsou texty shromažďovány v ohromných korpusech.

„Tím pádem jsou pro analýzu lexikální struktury vhodné texty jako například delší novinové články, krátké povídky nebo kapitoly románu (ovšem ne celý román). Jinými slovy přehnaně velké texty nejsou homogenní; jsou sémanticky členěny a sestávají se ze ‚sub-textů'."[45] Tuto vlastnost lze na zkoumaných datech vypozorovat, a jak Hřebíček dodává, konceptu homogenity dat jako první v dané problematice využil Altmann v roce 1992.[46]

„Je již zavedenou lingvistickou tradicí vytvářet povětšinou abstraktní lexikální významy slovních jednotek v jazyce tím, že se rozšíří prameny, z nichž se získávají lexikografická data. (…) Je pravidlem, že jazyková fakta jsou tím zkreslena."[47] Jedním ze vznikajících problémů je nahodilý předpoklad, že frekvence výskytu významu spojovaného s lexikální jednotkou je úměrný počtu významů spojovaných s touto lexikální jednotkou. „Odlišné významy jsou však, stejně jako drobné rozdíly lexikálních významů, zřetelné pouze v textech, ve kterých přímo probíhá přecházení na nové významy starých slov. Nelze spolehlivě rozlišovat mezi novými významy a záchvěvy starých významů." Na obě tyto kategorie je brán ohled v souvislosti s umístěním lexikální jednotky v mřížce tvořené textem, ale jejich vlastnosti se vytratí, když se zkoumaný korpus bezmezně rozroste.[48]

---

[44] HŘEBÍČEK, Luděk. Zipf's law and text. In *Glottometrics 3, 2002*, str. 28.

[45] Tamtéž, str. 29.

[46] Tamtéž.

[47] Tamtéž.

[48] Tamtéž.

Závěrem Hřebíčkova srovnání použitelných korpusů a textů pro analýzu na základě Zipfových zákonů samozřejmě není to, že by korpusy (ze své podstaty ve většině případů velmi rozsáhlé soubory jazykových dat) nebyly pro takovou analýzu vhodné vůbec. Jen je potřeba s nimi pracovat skutečně opatrně a podle konkrétního badatelského záměru pracovat s jejich velikostí – tedy nejlépe si ji pro své účely upravovat do verze ve zmenšené velikosti. Obecně zaměřené korpusy jsou tak podle něj sice užitečné co do praktického využití, teoretickou argumentaci však na jejich datech vystavět nelze.[49]

Luděk Hřebíček ve své práci nepracuje čistě jen se Zipfovým zákonem týkajícím se vztahu mezi rankem a frekvencí slov ve vybraném textu. Na téže turecké povídce[50] porovnává výsledky získané z analýzy pomocí Menzerath-Altmannova zákona, Mandelbrotova upřesnění Zipfova zákona, úpravy téhož zákona nazývající se Zipfova-Alekseevova distribuce. Pro potřeby této práce nejsou dílčí výsledky z jednotlivých analýz natolik důležité. Hřebíček se však na jejich pozadí věnuje i tématu interpretace získaných dat a opatrné práce s přiřazováním ranku. Pokud přistupujeme k analýze s výše zmíněným předpokladem, že každý text upřesňuje skutečný význam svých lexikálních jednotek, pak lze tvrdit, že k popsání těchto významů není tolik důležité, zda výzkumník pracuje s absolutní či relativní frekvencí slov, klíčové je pracovat s variabilním rankem. „Pokud má podmnožina lexikálních jednotek stejnou hodnotu frekvence, není možné, aby byly jednotky charakterizovány rozdílnou hodnotou ranku a jejich skutečné pořadí tak zůstalo skryté."[51] Ve snaze odstranit opakující se hodnoty frekvence přiřazené rozdílným lexikálním jednotkám se Hřebíček pokouší znovu analyzovat rozložení frekvencí slov s různými způsoby přiřazování ranku totožným frekvencím. „Zipfova původní myšlenka, tak jak je chápána, se nezabývá lexikálními jednotkami, ale jejich frekvencemi."[52] Hřebíček dle vlastních slov právě touto svou prací dokazuje, že i ranky samy o sobě mají svůj lingvistický smysl.[53]

---

[49] HŘEBÍČEK, Luděk. Zipf's law and text. In *Glottometrics 3, 2002*, str. 29.

[50] Povídka je dlouhá 721 slov a její slovník se sestává z 353 lexikálních jednotek. Pro potřeby analýzy ranku byly tyto jednotky uspořádány do sekvencí podle sestupné relativní frekvence.

[51] HŘEBÍČEK, Luděk. Zipf's law and text. In *Glottometrics 3, 2002*, str. 33.

[52] Tamtéž, str. 34.

[53] Tamtéž, str. 37.

Na závěr je třeba se vrátit k otázce, s jakými daty je nejvýhodnější pracovat, aby se na nich mohl nejlépe ověřovat Zipfův zákon vztahu mezi rankem a frekvencí. „Žádné lingvistické potvrzení platnosti Zipfova zákona nepotřebuje obrovské korpusy či texty, protože ty jsou vždycky statisticky nehomogenní. Jejich nezbytná nehomogenita stírá možnost vysledovat vlastnosti, které tento zákon vyjadřuje. Ty však mohou být nalezeny v samostatných textech, které nabízejí homogenní data. V textech se Zipfův zákon (především ve formulaci B. B. Mandelbrota) ukazuje být lingvisticky přesvědčivější než ve velkých korpusech či rozsáhlých textech. Těchto výsledků bylo dosaženo, když bylo v textech testováno rozložení frekvencí a lexikálních jednotek."[54]

3.2 Texty podle Zipfa

„Přirozené jazyky jsou komplexní systémy, které se vyvinuly jako efektivní dynamické struktury, které jsou schopné kódovat a přenášet velmi složité informace. (…) I když se vyvíjely v kratším časovém období než genetický kód a byly vystaveny rušnějšímu prostředí, své momentální komplexity také dosáhly pomocí přirozeného procesu vývoje."[55] Damián H. Zanette a Marcelo A. Montemurro zveřejnili v roce 2005 v periodiku *Journal of quantitative linguistic* společnou práci týkající se Zipfova rozložení v jazyce, ovšem nevydali se cestou, na níž by se původní zákon snažili ověřit či vyvrátit na vybraných textech. Svůj výzkum postavili na vytvoření dynamického stochastického modelu generujícího texty. „Model v sobě zahrnuje jak rysy související s obecnou strukturou jazyka, tak paměťový vliv, který je neoddělitelně spjatý s vytvářením dlouhých koherentních sdělení v komunikačním procesu."[56] Jejich cílem je zjistit, zda tento model, který respektuje základní teorii matematické lingvistiky, bude odpovídat přirozenému lidskému jazyku.

„Když je jazyk studován za pomoci nástrojů původně vyvinutých v rámci statistických zákonitostí, vynoří se velmi bohatá struktura na úrovních pohybujících se mezi skládáním jednoduchých slov za sebe a velkými organizačními vzorci o rozsahu

[54] Tamtéž.

[55] ZANETTE, Damián H., MONTEMURRO, Marcelo A. Dynamics of text generation with realistic Zipf distribution. In *Journal of quantitative linguistics 12, 2005*, str. 1.

[56] Tamtéž.

mnoha tisíc slov. Aplikování statistiky na záznamy lidského jazyka načrtává obraz jeho makroskopické struktury na úrovni popisu, který může odhalit stopy historie jeho vývoje a poskytnout informace o komplexním procesu, který probíhá na pozadí vytváření jazyka v mozku."[57] Jednou z nejobecnějších statistických vlastností psaného jazyka je podle autorů uznávaný Zipfův zákon „ve formě matematického poměru mezi rankem každého slova v soupise všech slov užitých v textu seřazených podle klesající frekvence a jejich frekvencí".[58]

I po padesáti letech od objevení pravidelností v jazyce, jež Zipfův zákon popisuje, je podle Montemurra a Zenetta samotný původ tohoto jevu těžko postižitelný. „Avšak pozoruhodná všudypřítomnost toho zákona u nejrůznějších jazyků naznačuje, že jeho původ musíme hledat ve velmi obecných pravděpodobnostních aspektech souvisejících s procesem vytváření jazyka."[59] Autoři v tomto kontextu neopomenou zmínit nedávné zjištění, že by Zipfův zákon mohl být výsledkem Markovova procesu skrytého v dynamice tvorby textu. Dále s touto tezí však nepracují z jednoduchého důvodu – jejích výsledků bylo dosaženo analyzováním biblických textů a Bible sama je spíše kolekcí relativně krátkých textů, proto u ní podle autorů může docházet ke statistickým výkyvům. Proto Zanette a Montemurro staví své další analyzování textu na pracích představených v kapitole *Přímé reakce na Zipfovy zákony*. „Za zmínku stojí dva modely vysvětlení Zipfova zákona, které reprezentují dva velmi odlišné postoje s respektem k lingvistickým důležitostem onoho zákona."[60] Jako první vyzdvihují práci Herberta Alexandra Simona, v níž přistupuje k tvorbě slovní zásoby jedince jako ke stochastickému procesu. „Simuluje dynamiku tvorby textu jakožto multiplikačního procesu, který u dlouhých textů vede asymptoticky k Zipfovu zákonu."[61] Druhá zmíněná práce je Mandelbrotova, která vysvětluje Zipfův zákon jako „neměnný rys statistické struktury náhodného sledu symbolů. Zatímco Simonův model propůjčuje

[57] ZANETTE, Damián H., MONTEMURRO, Marcelo A. Dynamics of text generation with realistic Zipf distribution. In *Journal of quantitative linguistics 12, 2005*, str. 1.

[58] Tamtéž.

[59] Tamtéž.

[60] ZANETTE, Damián H., MONTEMURRO, Marcelo A. Dynamics of text generation with realistic Zipf distribution. In *Journal of quantitative linguistics 12, 2005*, str. 1.

[61] ZANETTE, Damián H., MONTEMURRO, Marcelo A. Dynamics of text generation with realistic Zipf distribution. In *Journal of quantitative linguistics 12, 2005*, str. 1.

Zipfovu zákonu nikoli nedůležitou lingvistickou významnost, Mandelbrotovo vysvětlení podává zákon jako pouhou vlastnost náhodného souboru vlastností."[62]

Zanette a Montemurro jsou přesvědčeni, že jejich práce pomůže vyřešit střet mezi těmito pohledy vysvětlením konkrétních odchylek, které se projeví při empirické distribuci. Ve své práci představují dynamický model vysvětlující empirické chování frekvencí slov jako následek dvou procesů, jimiž jsou „souhrnný paměťový efekt řízený kontextem, který je v zásadě spojen se vzájemnou interakcí mezi multiplikačními a aditivními dynamikami ve výběru slov a lokální, na gramatice nezávislý efekt, který je spojovaný se vznikem flektivních slov. Tento model je schopen napodobit realistické Zipfovo rozložení a krom toho obsahuje i lingvistický výklad."[63]

Zanette a Montemurro museli pro své potřeby trochu upravit modely Zipfova zákona, které představili jako základ své práce. „Smyslem Simonova modelu je zachytit základní rysy skutečného generování textu tím, že specifikuje, jak jsou slova přidávána do textu: předpokládejme, že v každém kroku *t* je přidáno nové slovo do textu, který začíná pouze jedním slovem v kroku $t = 1$, tím pádem je v jakémkoli kroku délka textu rovna hodnotě *t*. S danou pravděpodobností α je nové slovo, které se v textu dosud neobjevilo, přidáno v $t + 1$, nebo s komplementární pravděpodobností je nové slovo vybráno náhodně mezi předchozími *t* slovy."[64] Jelikož je pravděpodobnost opakování slov, která se již v textu objevila, úměrná počtu jejich předchozích výskytů, znamená to podle autorů, že se v takto fungujícím postupu buduje silná konkurence mezi různými slovy.[65]

„Srovnání s distribucí získané ze skutečných textů odhaluje kvalitativní změnu, třebaže jsou současně patrné kvantitativní rozdíly. Konkrétně Simonův model nenapodobuje rychlejší rozpad v nižších frekvencích (…). Původní dynamika Simonova modelu v sobě zahrnuje skutečnost, že když musí být slovo vybráno z textu, který byl prozatím napsán, pravděpodobnost, že bude slovo zopakováno, je úměrná počtu jeho předchozího výskytu."[66] Autoři jsou však přesvědčeni, že nově přidané slovo nemá

[62] ZANETTE, Damián H., MONTEMURRO, Marcelo A. Dynamics of text generation with realistic Zipf distribution. In *Journal of quantitative linguistics 12, 2005*, str. 1.

[63] Tamtéž.

[64] Tamtéž, str. 2.

[65] Tamtéž.

[66] ZANETTE, Damián H., MONTEMURRO, Marcelo A. Dynamics of text generation with realistic Zipf distribution. In *Journal of quantitative linguistics 12, 2005*, str. 2.

zpočátku zřetelně vymezený vliv na kontext celého textu, a proto by bylo záhodno pracovat s pravděpodobností jeho dalšího výskytu trochu jinak a opatrněji. Pro zjednodušení oné operace si k úpravě Simonova modelu vybrali exponenciální distribuci, pro niž museli přesně stanovit pouze jeden vymezující faktor. Tím byl průměrný práh závislosti na slovu (word-dependent treshold).[67]

Analýza byla provedena na literárních dílech, konkrétně se jejich modelu podařilo zreprodukovat distribuci odpovídající té podle Zipfových zákonů jak u textů psaných v jazycích s bohatou flexí (kupříkladu ruština, latina), tak u textů neflektivních jazyků jako angličtina či španělština. „Ve všech případech bylo dosaženo rychlejšího rozpadu, projevujícího se v distribuci vztahu rank-frekvence u vyšších ranků, zvolením parametru prahu závislosti na slově (word-dependent treshold) mezi 2 a 4."[68] Tento parametr bohužel nemůže být do Simonova modelu importován přímo, autoři se však snaží přiblížit jeho rovnicím tím, že proces ještě o něco více zjednodušují dalšími úpravami a dostavují se jim výsledky. „Abychom to shrnuli, tak jsme se zkoumali původ Zipfova zákona, který je považován za nejzákladnější statistický vzorec nalezený v psaném lidském jazyce. Naše výsledky potvrzují, že distribuce vztahu ranku a frekvence slov je především důsledkem multiplikačních procesů skrytých pod povrchem průběhu generování jazyka. Kromě toho jsme byli schopni propojit drobné detaily empirické distribuce s dvěma klíčovými subprocesy, které se podílejí na tvorbě textu."[69] Tím je jednak pravidlo nárůstu slovníku, které do sebe začlenilo i vliv struktury jazyků nezaložených na flexi, jednak dynamika nově se objevujících slov při vytváření textu.

Za svůj hlavní závěr autoři považují to, že jejich "lingvisticky citlivá poupravení" [70] modelu, původně představeného Herbertem Alexandrem Simonem, odstraňují drobné odchylky vzniknuvší mezi výsledky Simonova modelu a skutečné distribuce kopírující Zipfovy zákony, čímž přispívají k vnímání těchto zákonů jako lingvisticky významného rysu psaných textů.

---

[67] ZANETTE, Damián H., MONTEMURRO, Marcelo A. Dynamics of text generation with realistic Zipf distribution. In *Journal of quantitative linguistics 12, 2005*, str. 2.

[68] Tamtéž, str. 3.

[69] Tamtéž, str. 4.

[70] Tamtéž.

3.3 Nevyhnutelnost Zipfovy distribuce v komunikaci

George Kingley Zipf se ve své práci pojednávající o principu nejmenšího úsilí dotkl také určitého napětí vznikajícího z protichůdné realizace onoho principu u účastníků komunikace. Pro mluvčího je z hlediska principu nejmenšího úsilí výhodné vytvářet svá sdělení z co nejméně obsáhlé slovní zásoby, aby nemusel vynakládat námahu na udržení rozsáhlého slovníku v paměti a posléze věnovat zvýšenou pozornost opatrnému a pečlivému výběru jednotlivých slov tak, aby nejlépe sloužila jeho záměru a dokázala beze zbytku vyjádřit přesně to, co měl na mysli, aniž by docházelo k větším odchylkám ve sdíleném významu a tím i chybám v komunikačním procesu sdíleném s příjemcem. „Z pohledu mluvčího (*ekonomizace mluvčího*), který má za úkol výběr nejen významů ve sdělení, ale také slov, jimiž tyto významy vyjádří, by důležitá skrytá ekonomizace bezpochyby spočívala ve slovní zásobě čítající výhradně jedno slovo – jediné slovo, které by znamenalo, cokoli by mluvčí chtěl, aby znamenalo. Tudíž pokud by existovalo *m* různých významů, které by se mluvčí snažil vyjádřit slovy, samo toto jediné slovo by mělo oněch *m* různých významů.“[71]

Avšak taková potřeba ekonomizace vyjadřování u mluvčího zcela naráží na druh jazykové ekonomizace, kterou by naopak přivítal příjemce sdělení. Pokud by byl nucen pracovat s vyjádřením vytvořeným pomocí slovní zásoby obsahující výlučně jen jediné slovo, pak by ho stálo velké úsilí přijít na to, který ze všech možných významů právě v tuto chvíli ono jediné slovo v různém postavení nese. Podle Zipfových slov by přímo „čelil nemožnému úkolu“. [72] „Z pohledu posluchače (*ekonomizace posluchače*), jehož úkolem je dešifrovat významy zvolené mluvčím, by důležitou vnitřní ekonomizací řeči ve skutečnosti byla spíše slovní zásoba takového rozsahu, který by obsáhl výrazně rozdílná slova, každé pro odlišný význam, jenž je slovy vyjádřen. Tudíž pokud by existovalo *m* různých významů, existovalo by *m* různých slov, každé o jednom jediném významu.“[73] Čímž by se posluchači silně ulehčila práce, neboť by byl ušetřen námahy, kterou by ho stálo vybírání předpokládaně nejvhodnějšího významu u slov z velmi omezeného slovníku, jakým ten, který je vyžadován ekonomizací mluvčího, zajisté je.

[71] ZIPF, George. *Human behavior and the principle of least effort.* 1949, str. 20.

[72] Tamtéž, str. 21.

[73] Tamtéž.

Tyto dvě úsporné tendence jdou svou povahou samozřejmě zcela proti sobě. George Kingsley Zipf je ve své práci nazývá *protichůdnými silami.*[74] „Jedna ‚síla' (*ekonomizace mluvčího*) tíhne k zredukování velikosti slovní zásoby na jediné slovo tím, že pod tímto slovem sjednotí všechny významy; z toho důvodu ji můžeme náležitě nazývat *síla unifikace*. V protikladu k síle unifikace je druhá ‚síla' (*ekonomizace posluchače*), která má sklon k rozšiřování velikosti slovní zásoby až do bodu, kdy bude existovat zřetelně odlišné slovo pro každý zřetelně odlišný význam. Jelikož tato druhá ‚síla' směřuje k rozšíření rozmanitosti slovní zásoby, budeme jí od nynějška nazývat *síla diverzifikace*."[75]

A právě na tyto opozitní ekonomizační tendence a z nich vyplývající možnosti matematicky měřit jejich důsledky v psaných textech poukazují v úvodu svého článku autoři Bernat Corominas-Murtra, Jordi Fortuny Andreu a Ricard Vincente Solé. Jejich společný text byl publikován roku 2011 a zabývají se v něm tím, jak se ve vývoji komunikace objevují pravidelnosti popsané Gergem Kingsley Zipfem. „Za použití formalismu pevně zakořeněného v struktuře teorie informace názorně předvedeme, že Zipfův zákon je jediný předpokládaný výsledek vyvíjejícího se komunikativního systému v souladu s nekompromisní definicí komunikačního napětí popsaného Zipfem." [76] Komunikačním napětím jsou míněny výše popsané protichůdné ekonomizační síly mluvčího a posluchače.

„Nedávný přístup, dalece přesahující komunikační rámec, definuje spletité charakteristiky systému klíčové k představení statistických případů následujících Zipfův zákon: otevřené, neomezené množství dosažitelných stavů a lineární úbytek entropie v důsledku standardních vnitřních omezení. Onomu lineárnímu úbytku entropie se daří uchopit intuitivní představu, že studovaný systém je v *přechodovém stavu* mezi uspořádáním a neuspořádáním, nebo že možné informativní napětí je vyrovnáno a neohraničené množství dosažitelných stavů odráží jejich přístupnou podstatu. Ukázalo se, že za velmi všeobecné parametrizace a při zakomponování škálové invariance

[74] ZIPF, George. *Human behavior and the principle of least effort.* 1949, str. 21.

[75] Tamtéž.

[76] COROMINAS-MURTA, Bernat, FORTUNY ANDREU, Jordi, SOLÉ, Ricard Vincente. Emergence of Zipf's law in the evolution of communication. In *Physical review E 83, 2011*, str. 1.

do řešení, je Zipfův zákon jediný možný výsledek."[77] Kromě toho, že autoři zasazují rozpor mezi ekonomizací komunikace mluvčího a ekonomizací komunikace posluchače do matematického formalizačního rámce, pokoušejí se zformalizovat i vývojové tendence v komunikaci a tvorbě kódu. I jiné rostoucí systémy se nacházejí v určité rovnováze, stejně jako komunikace probíhající mezi mluvčím a posluchačem – jazykový systém se neustále rozrůstá, ale současně se nachází ve vcelku rovnovážném stavu, neboť kdyby tomu tak nebylo, mluvčí a posluchače bude stát nemalé úsilí si navzájem porozumět. Proto chtějí Corominas-Murta, Fortuny Andreau a Solé vzít ve své práci v potaz vývoj komunikační směny na pozadí vzrůstajícího systému. „Vývojová složka je variačně představována pomocí snižování rozdílnosti na minimum mezi konfiguračními kódy náležejícími úspěšným krokům v čase vývoje. Tyto minimální změny vycházejí z takzvaného *Principu nejmenšího informačního rozlišení (Minimum discrimination information principle*, proto označováno i jako *M D I P*), obecného variačního principu analogického principu maximální entropie."[78]

Princip minimální diskriminace informace udává, že pokud nastane situace, při níž se změní vymezení systému, nejočekávanější distribuce je s největší pravděpodobností ta, jež se snaží na nejnižší míru ubírat z Kullback-Leiblerovy divergence, nazývané též diskriminační informace (tento název podporoval především Kullack). Diskriminační informace je pojmenování pro takové množství informace, které zpráva poskytuje pro odlišení pravděpodobnosti její distribuce. Autoři textu se dle vlastních slov pokusí za použití M D I P poskytnout důkaz o jednoznačnosti Zipfových zákonů a jejich vznik ve vyvíjejících se kódech. Nejprve však věnují svou pozornost komunikačnímu napětí mezi mluvčím a posluchačem (popř. kodérem a dekodérem), které se snaží matematicky formulovat a po převedení do matematického rámce také ideálně vyřešit. Ustanovují formulaci obsahující v sobě symetricky vyváženou rovnováhu mezi ekonomizační snahou mluvčího (kodéra) a ekonomizační snahou posluchače (dekodéra). Tuto formuli nazývají *podmínkou rovnovážnosti* a vyzdvihují její dosažení pomocí kooperačních principů mezi předloženými požadavky

[77] COROMINAS-MURTA, Bernat, FORTUNY ANDREU, Jordi, SOLÉ, Ricard Vincente. Emergence of Zipf's law in the evolution of communication. In *Physical review E 83, 2011*, str. 1–2.

[78] Tamtéž, str. 2.

mluvčího (kodéra) a posluchače (dekodéra). Avšak ani toto zjištění samo o sobě nestačí na objasnění toho, že se Zipfův zákon objevuje při vývoji komunikace a jazyka.

Corominas-Mutra, Fortuny Andreu a Solé a vycházejí při dalším řešení z úvodního předpokladu, že s postupujícím časem se zvyšují naše schopnosti obstát či přímo uspět v komunikaci, čímž se zároveň zvyšuje počet vstupních znaků, jež jsme schopni do našeho komunikačního výstupu zakódovat. Tím se ovšem zvyšuje i entropie vstupních sad, jejichž velikost se může stát neomezenou a sady se začínají chovat jako ohromná zásobárna informací. „Pokud potenciální informační bohatství vstupní sady je neomezené, potom je informační i bohatství výstupní sady, tedy za určitých omezení způsobených a vnucených podmínkou rovnovážnosti."[79]

Když se autoři po předchozích krocích konečně dostávají k původně předestřenému jádru své práce, tedy objasnění výskytu Zipfovy distribuce v jazyce během vývoje kódu, plně využijí dříve představený Princip nejmenšího informačního rozlišení. Ten podle nich dokáže přímo provést vyvinutí kódu, který musí na své vývojové cestě uspokojit všechny v něm zainteresované strany. „Rozhodující přínos Principu nejmenšího informačního rozlišení je v tom, že přirozeně zavádí stopy závislosti vnucené vývojem. Následuje termodynamickou metaforu, čímž se tento variační princip v našem kontextu chová jako pravidlo minimalizace energie působící v průběhu přenosů úspěšných kódů."[80]

Další, velmi zajímavý bod je těsně svázán s kódem, jímž se projevuje Zipfův zákon, konkrétněji s důsledky symetrické okrajové podmínky (symmetry condition), matematického přístupu, který abstraktně kóduje Zipfovu hypotézu slovníkové vyváženosti: přítomnost nevyhnutelné nejednoznačnosti v kódu. Je běžné přijímáno, že přirozené jazyky jsou nejednoznačné, konkrétně že lingvistické výroky mohou mít více než jednu interpretaci. Funguje-li princip nejmenšího úsilí a mezi kodérem a dekodérem je tedy kooperativní strategie, přítomnost určité míry nejednoznačnosti může být očekávána, pokud má mluvčí tendenci přidělovat některým signálům více než jeden význam. Nejednoznačnost je tedy vedlejším produktem efektivní komunikace spíše než znakem špatného komunikačního plánu.

---

[79] COROMINAS-MURTA, Bernat, FORTUNY ANDREU, Jordi, SOLÉ, Ricard Vincente. Emergence of Zipf's law in the evolution of communication. In *Physical review E 83, 2011*, str. 3.

[80] Tamtéž, str. 4.

Zipfův zákon se objevuje v systému, kde se kodér a dekodér vyvíjí spolu s obecným problémem minimalizace energie. Rozsah aplikovatelnosti na fenomény v reálném světě však musí být porovnán s platností dat, bylo totiž poukázáno na fakt, že mnoho předpokládaných chování, které vykazují mocninný zákon, ukazuje odchylky, když je statistická analýza provedena přesně. Je však nutné podotknout, že odchylka od předpokládaného chování nemusí nutně být přisouzena selhání rámce. Je třeba vzít v úvahu, že další omezení jako například obecná omezení paměti mohou ovlivnit konečné rozložení.

4 Zipfovy zákony a slova

Tak jako v minulé kapitole stály texty tematicky spíše vedle sebe, v této části se texty mezi sebou prolínají, někteří autoři docházejí nad toutéž problematikou i ke zcela protichůdným závěrům. Tak je tomu například v otázce toho, zda je možno rozšířit měřitelné výsledky Zipfových zákonů ze slov i na fráze nebo jestli to možné není a zda je pro jakékoli stanovení výsledku třeba pracovat pouze se slovy. A když už s nimi pracujeme, jakou délku by toto slovo mělo mít?

Problematické je již určení slova v textu, neboť slovo je možno vymezit mnoha způsoby. „Povaha slov je určována v indoevropských jazycích nejvíce jejich syntaktickou funkcí ve větě jako determinantů. Tím je také dána definice slova ve smyslu syntaktickém: je to znak, jehož žádné determinans nelze blíže určit. Od slova syntaktického nutno rozeznávat slovo sémantické, které je jazykovým znakem nebo kombinací jazykových znaků, které označují něco v mimojazykové realitě.“[81] Sémantické slovo však neodpovídá vždy přesně slovu syntaktickému. Stejně jako se nemusí vždy krýt slovo sémantické a fonologické, které je jednotkou projevující se fonologickými prostředky, jimiž je možno vyjádřit jádro sémantického slova. Fonetické slovo se vyznačuje fonetickými prostředky, grafické slovo je určováno grafickými prostředky spojenými do jednoho celku. „V jazyce možno tedy pod názvem ‚slovo’ rozumět různé věci podle toho, v jaké vrstvě jazyka jsme. Jisto je, že základem všech těchto různých pojetí slova je vždy přímo nebo nepřímo slovo sémantické.

[81] LYER, Stanislav. Slovo a jeho struktura. In *Slovo a slovesnost r. 8,* č. *1, 1942.*

Slovo syntaktické, fonologické apod. jsou vlastně jen různé realizace sémantického slova v různých vrstvách jazyka. Na základě této úvahy možno tedy slovo definovat jako samostatnou jednotku, která se projevuje použitím prostředků, jimiž se v dané vrstvě jazykového systému vyjadřuje přímo nebo nepřímo jednota sémantického slova."[82]

4.1 Délka slova

Analýza délky slova velmi úzce souvisí právě se samotným určením, co lze považovat za slovo. „Problematika týkající se analýzy délky slova stojí v centru pozornosti jak lingvistů, tak matematiků již velmi dlouhou dobu. Navzdory všemu úsilí rozsah problémů asociovaných s problematikou délky slova stále roste: na jedné straně kvantitativní analýzy dalších a dalších jazyků přinášejí nové poznatky o charakteru frekvenční distribuce slov obecně, na straně druhé se ukazuje, že mnoho nově definovaných vlastností jazyka úzce souvisí právě s délkou slova."[83] K té lze od začátku přistupovat jako k fenoménu, jehož rozložení v textu nijak nepodléhá chaotickému chování, řídí se tedy určitými zákony. Kromě toho se v rámci jednoho jazyka liší jak zmíněná distribuce délky, tak například i průměr délky slova a to v závislosti na typu textu, žánru, stylu, autorství a podobně.[84]

„Původní Zipfův zákon se zabývá statistikou orankovaných slov v přirozených jazycích. Před nedávnem bylo toto zevšeobecněno na ‚slova' definovaná jako n-tice symbolů odvozených překladem reálných hodnot jednorozměrných časových řad do doslovných posloupností."[85] Amir Hossein Darooneh a Rahmani publikovali v roce 2009 článek, v němž se zabývají právě délkou slov, která jsou předmětem analýzy vztahu ranku a frekvence. Tak jako bylo třeba na začátku kapitoly zmínit, že uchopit slovo pro jeho následnou analýzu není nikterak jednoduchý úkol, totéž platí pro práci, v níž si autoři sami kladou za cíl pracovat s délkou slov podrobovaných

[82] LYER, Stanislav. Slovo a jeho struktura. In *Slovo a slovesnost r. 8,* č*. 1, 1942*

[83] ČECH, Radek, POPESCU, Ioan-Iovitz, ALTMANN, Gabriel. *Metody kvantitativní analýzy (nejen) básnických textů.* 2014, str. 75.

[84] Tamtéž.

[85] DAROONEH, A.H., RAHMANI, B. Finite size correction for fixed word length Zipf analysis. In *The European physical journal B 70, 2009*, str. 287.

matematicko-lingvistickému zkoumání. Začínají osvětlením termínů z úvodní všeobecné definice slova: „Časové řady jsou definovány jako odběr vzorků v průběhu času v rovnoměrných vzdálenostech.“ [86] Jinými slovy jde o chronologicky zorganizovanou posloupnost naměřených hodnot předem zvoleného statistického indikátoru, přičemž toto časové uspořádání je uskutečnění náhodného děje.

Pokud se ve vývoji dvou časových řad objeví určitá míra závislosti, pak se jedná o korelaci. A analýza stojící na podkladech Zipfovy analýzy se ukázala jako snadno uchopitelný způsob kvantifikace korelací časových řad.[87] „Tato technika je založena na překládání často se opakujících časových řad do posloupnosti symbolů a počítání frekvencí každého slova, to znamená vzorce po sobě jdoucích symbolů. Přiřazením ranků těmto slovům podle jejich frekvence od nejčastějšího po nejméně obvyklé a zakreslením do grafu logaritmů frekvence oproti logaritmům ranku nám vytvoří grafický zápis Zipfa.[88] Pro ten nejnižší rank se bod použitý na vykreslení pozice v grafu jeví jako by po přímce padal.“[89] Darooneh a Rahmani se pokusili modifikovat Zipfův zákon do takové podoby, aby bral ideálně v úvahu dopad konečné velikosti vzorku, což by mělo zlepšit jeho schopnost odhadu a umožnit výzkumníkům používat jej jako charakteristickou hodnotu pro dlouhodobé korelace v časových řadách. „Abychom potvrdili toto prohlášení, použijeme jak Zipfovu analýzu, tak metodu DFA pro kvantifikování korelací v časových řadách a porovnáme výsledky jejich odpovídajících charakteristik.“[90]

DFA znamená *Detrended fluctuation analysis*, jejíž podstatou je „rozsekat signál na intervaly a v každém intervalu proložit signál přímkou a spočítat odchylku od této přímky. Tento postup se pak provede pro různě velké intervaly neboli box size. Následně se vykreslí délka těchto intervalů na ose x proti celkové odchylce od přímek na ose y, ovšem na logaritmické škále tedy log (box size) proti log (total DFA).

[86] DAROONEH, A.H., RAHMANI, B. Finite size correction for fixed word length Zipf analysis. In *The European physical journal B 70, 2009*, str. 287.

[87] Tamtéž.

[88] Míněn 45° úhel, který je zachycen v úvodní kapitole při popisu Zipfových zákonů.

[89] DAROONEH, A.H., RAHMANI, B. Finite size correction for fixed word length Zipf analysis. In *The European physical journal B 70, 2009*, str. 287.

[90] Tamtéž.

Získaný graf má směrnici, která se nazývá celková DFA."[91] Při použití této metody je sice zapotřebí analyzovat delší časové období, její výhoda však beze zbytku spočívá v tom, že je s její pomocí možno odlišit, které zakolísání v předpokládané pravidelnosti dat je způsobeno vnějšími vlivy a které naopak svou existencí zrcadlí vnitřní dynamiku systému.[92]

Darooneh a Rahmani se ve svém textu vracejí k Zipfovu zákonu a analýze, která jej stvrzuje a stojí na přiřazování ranku slovům přirozeného jazyka dle jejich frekvence v určitém předem daném korpusu. „Body vykreslené v grafu náležící nejnižším rankům se nacházejí okolo přímé linie. Absolutní hodnota, která se nejlépe slučuje se svažující se přímkou, je nazývána Zipfovým exponentem." [93] Uvedené rozložení autoři jasně demonstrují na výsledcích analýzy textů z perské literatury, jejichž zanesení do grafu skutečně zcela dokonale kopíruje přímku svažující se zleva doprava pod pětačtyřicetistupňovým úhlem.

Ovšem jak už bylo řečeno, většina výsledků distribuce vztahu ranku a frekvence podle Zipfova zákona se v grafu nevykreslí podél nenarušené přímky, ale vykazují u jejího spodního konce odchylky v podobě strmějšího pádu po ose y. „Má se za to, že ohraničená velikost dat je tím hlavním důvodem pro překotně zrychlené klesání na konci Zipfova grafu. Tato skutečnost zapříčiňuje, že postup měření Zipfova exponentu je nejednoznačný." [94] A exponent autoři potřebují přesnější, protože jejich záměrem je použít jej pro kvantifikaci korelací v časových řadách. Proto přistupují k úpravě výpočtů Zipfovy distribuce ať už na úrovni zpracování hrubých dat, nebo propočtem týchž měření pomocí *Detrended fluctuation analysis*. Zajímavá je však především jejich práce s relativní frekvencí, jejímž zavedením se autoři pokouší zjemnit strmý pád na konci přímky v grafu s vykreslenou Zipfovou distribucí. „Relativní frekvence slova je vypočítána vydělením jeho patrné frekvence součinem frekvencí všech znaků onoho slova." [95] Podle Darooneha a Rahmaniho

---

[91] ŽOUŽELKOVÁ, Jana. *Nové míry volatility ekonomických časových řad.* Olomouc, 2012. Diplomová práce. Univerzita Palackého v Olomouci. Přírodovědecká fakulta. Katedra matematické analýzy a aplikací matematiky. Vedoucí práce: RNDr. Tomáš Fürst, Ph.D., str. 89.

[92] Tamtéž.

[93] DAROONEH, A.H., RAHMANI, B. Finite size correction for fixed word length Zipf analysis. In *The European physical journal B 70, 2009*, str. 288.

[94] DAROONEH, A.H., RAHMANI, B. Finite size correction for fixed word length Zipf analysis. In *The European physical journal B 70, 2009*, str. 288-289.

[95] Tamtéž, str. 289.

je tak možné vypočítat Zipfův exponent pro jakákoli data, pokud se ve vztahu k ranku bere v úvahu relativní frekvence. Po takto upraveném výpočtu se přímka tvořená získanými daty v grafu o něco srovná.

Proto je důležité téma definice slova i délky analyzovaných slov. Pokud by totiž i další autoři chtěli dosáhnout grafu Zipfovy distribuce ve tvaru, který kopíruje tvar přímky bez strmějšího klesání po ose y a použili by k tomu matematické úpravy vybrané Amirem Hosseinem Daroonehem a jeho kolegou Rahmanim, pak by jejich výsledky velmi závisely na prvotním vydefinování, co z textu bude považováno za slovo. Z přístupu ke slovu se totiž bude odvíjet i jeho délka, a protože způsob Darooneha a Rahmaniho ve svém postupu využívá výpočet relativní frekvence slova, jež se počítá i z frekvence každého jednotlivého znaku slova, pak délka každého slova ovlivňuje jeho relativní frekvenci a tím i autorskou distribuci vztahu ranku a frekvence podle původního Zipfova zákona.

4.2 Zipfova distribuce slov a frází

"Zipfův zákon konstatuje, že frekvence slovních tokenů ve velkém korpusu přirozeného jazyka je nepřímo úměrná jejich ranku." [96] Le Quan Ha, Elvira Sicilia-Garcia, Ji Ming a Jack Smith představili v lednu roku 2002 při příležitosti 19. ročníku Mezinárodní konference počítačové lingvistiky v Japonsku (International conference on computation linguistic) své společné pojednání, ve kterém platnost této základní premisy prvního Zipfova zákona ověřují na dvou rozsáhlých jazykových korpusech, nejprve na angličtině, poté na mandarínštině. Nepodrobují však analýze pouze slova, ale též n-gramové slovní fráze.

V úvodu práce připomínají, že Zipf objevil a zformuloval svůj zákon poté, co ručně prozkoumal modernistický román *Ulysses* [97] od irského spisovatele Jamese Joyce, který byl nejprve vydáván od března 1918 do prosince 1920 po částech jako seriál v americkém časopise The Little review a teprve o dva roky později vyšel knižně v pařížském nakladatelství. Slovník vytvořen z tého knihy čítal 29 899 různých

[96] HA, Le Quan, SICILIA-GARCIA, Elvira, MING, Ji, SMITH, Jack. *Extension of Zipf's Law to Words and Phrases,* 2002, str. 1.
[97] Ulysses je římská verze jména postavy řecké mytologie Odyssea, na jehož příběh je v knize mnohokrát odkazováno. Proto je v češtině kniha vydávána pod názvem, která odpovídá řecké variantě této bájné figuty - Odysseus.

slovních typů propojených s 260 430 slovními tokeny. Využití rozvoje počítačů v 60. letech přispělo k potvrzení, že zákon platí pro malé korpusy, které v té době mohly být zpracovány.[98] Sklony křivek vzešlých z tohoto potvrzování se následkem Mandelbrotových úprav zákona od sebe poněkud lišily, obecně se však mělo za to, že jde jen o drobné odchylky od původního zákona. Tedy s výjimkou právnických textů, které vykazovaly výrazné odchylky v konstantách, jež se nakonec ukázaly jak stvrzení dojmu, že právníci používají více slov než ostatní lidé.[99] „Zpracovávání větších korpusů s jedním milionem slov či více bylo usnadněno rozvojem osobních počítačů v 80. letech 20. století. Když byly narýsovány Zipfovy křivky pro tyto korpusy, ukázalo se, že na spodu klesnou pod Zipfovu přímou linii," což se začínalo projevovat u umístění slov s rankem vyšším než 5000.[100] Toto zjištění nahrávalo především oponentům Simonovy derivace, protože se tím zdálo nepochybné, že zákon zcela evidentně neobstojí u jazykových korpusů, v jejichž zpracovaných datech se může objevit rank vyššího čísla než je 5000. Ha, Sicilia-Garcia, Ming a Smith k tomuto závěru však nehodlají přistupovat jako k výchozí definici a sami jej raději na svém korpusovém materiálu ověřují.

Autoři začínají s angličtinou, jejíž korpus, s nímž pracují, je sestaven z čísel *Wall Street Journal* ze tří různých let o různé velikosti. V subkorpusu čísel z roku 1987 je zhruba 19 milionů tokenů, výtisky vybrané z roku 1988 jich v sobě mají 16 milionů a subkorpus sestaven z exemplářů vydaných roku 1989 zahrnuje 6 milionů tokenů. Křivky Zipfovy distribuce těchto korpusů vykreslené do grafu jsou souběžné a vykazují obdobnou strukturu. „Jazyk ale není tvořen jednotlivými slovy, nýbrž se sestává z ustálených slovních spojení ze dvou, tří a více slov, které jsou obvykle nazývány n-gramy pro n=2, 3 a tak dále. Pro každou hodnotu n mezi 2 a 5 jsme vypočítali frekvence všech n-gramů v každém korpusu a seřadili je podle ranku, jako jsme to učinili se slovy."[101] Tento krok autorům umožnil vykreslit do grafu křivky bigramů, trigramů, 4-gramů a 5-gramů, které následně porovnali i s křivkami slov. Nutno podotknout, že u všech tří oddělených korpusů se křivky n-gramů chovají naprosto totožně, jejich tvar je mezi korpusy bez problému zaměnitelný.

[98] HA, Le Quan, SICILIA-GARCIA, Elvira, MING, Ji, SMITH, Jack. *Extension of Zipf's Law to Words and Phrases,* 2002, str. 1.
[99] Tamtéž, str. 1-2.
[100] Tamtéž, str. 2.
[101] HA, Le Quan, SICILIA-GARCIA, Elvira, MING, Ji, SMITH, Jack. *Extension of Zipf's Law to Words and Phrases,* 2002, str. 2.

„Povšimněte si, že křivka unigramu kříží křivku bigramu, když se rank přibližně rovná 3000, a to ve všech třech případech."[102]

Pro analýzu jazyka mandarínské čínštiny autři využili korpusu TREC, který obsahuje články periodika *People's daily newspaper* vydané v čase mezi lednem roku 1991 a prosincem roku 1993 spolu s články z tiskové agentury *Xinhua news agency* zveřejňovanými mezi dubnem 1994 a zářím 1995. Celkově TREC disponuje databází 19 546 872 tokenů, což se velikostí velmi přibližuje největšímu korpusu použitému pro analýzu anglického jazyka. „Jazyk mandarínské čínštiny je slabičný jazyk, ve kterém každá slabika je ve stejnou chvíli i slovem a čínských znakem. Jink řečeno, složená slova jsou vytvářena kombinování slabik dohromady, podobně jako se v angličtině tvoří n-gramy."[103] Počet slabikových typů, což vlastně znamená unigramů, je v korpusu TREC pouze 6 300, což je oproti 114 718 slovních typů výrazně nižší počet, proto „není divu, že Zipfova křivka pro unigramy v mandarínské čínštině se velmi liší od Zipfovy křivky pro unigramy v angličtině. (…) Krom unigramů jsou si tvary dalších Zipfových křivek pro n-gramy z korpusu TREC podobné s těmi pro anglický jazyk, nejsou však úplně stejné. Konkrétně křivka bigramu pro mandarínskou čínštinu je mnohem více zakřivený než anglická křivka, protože v mandarínské čínštině je více složených slov než v angličtině."[104]

Ha, Sicilia-Garcia, Ming a Smith si kladou otázku, proč Simonova derivace původního Zipfova zákona opravdu selhává, jakmile je ve zkoumaném vzorku textů více než 5000 slovních typů a docházejí k závěru, že je to zřejmě vinou již zcela počátečního Simonova přistoupení k výpočtu pravděpodobnosti objevování se nových slov v analyzovaném textu. „V rozhodujících chvílích své derivace Simon přisoudil pravděpodobnost nově se objevujícímu slovu v korpusu jako by zavádělo nějaký nový význam, dosud nevyjádřený žádnými slovy. Nicméně jakmile počet slov narůstá, nové myšlenky jsou často vyjadřovány ne v jednotlivých slovech, ale ve frázích o několika slovech či za pomoci slov složených."[105] To však Herbert Alexandr Simon ve své práci navazující na tu Zipfovu vůbec nebral do úvahy. Pokud by tak ve své době učinil, pak by zákon jím upravený obsahoval nejen slova, ale též fráze. Proto se Ha,

[102] HA, Le Quan, SICILIA-GARCIA, Elvira, MING, Ji, SMITH, Jack. *Extension of Zipf's Law to Words and Phrases,* 2002, str. 2.
[103] Tamtéž, str. 3.
[104] Tamtéž, str. 4.
[105] HA, Le Quan, SICILIA-GARCIA, Elvira, MING, Ji, SMITH, Jack. *Extension of Zipf's Law to Words and Phrases,* 2002, str. 5.

Sicilia-Garcia, Ming a Smith domnívají, že by snad i původní zákon George Kingsleyho Zipfa měl zahrnovat slova stejně jako fráze. „Z toho důvodu jsme dali dohromady všechny ungramy a n-gramy i s jejich frekvencemi do jednoho velkého souboru roztříděného podle frekvencí a s přiřazeným pořadím ranku, jako jsme to učinili se všemi daty předtím.“[106] Takto propojená data se po vynesení do grafu velmi výrazně překrývají. Především pak ranky o vyšší hodnotě než 100 u zkombinovaných dat pro oba jazyky vykazují totožný sklon přímek v grafu. A n-gramy, pro něž platí n je větší či rovno 2, zase svými křivkami vynesenými v grafu odpovídají odchylkám různých unigramů též vzešlých ze Zipfova zákona.[107]

„Tyto výsledky se jeví jako prokázání správnosti Simonovy derivace. Nicméně, ať už je Simonova derivace zcela platná či nikoli, výsledky jsou novým potvrzením původního Zipfova zákona v rozšířeném tvaru. Tyhle pozororuhodné výsledky byly předběžnými experimenty shledány opodstatněnými ve třech dalších jazycích: irštině, latině a vietnamštině.“[108] Ha, Sicilia-Garcia, Ming a Smith svou práci na konferenci tedy uzavírají s tím, že znovu potvrdili platnost původního Zipfova zákona i na jazykových korpusech mnohem větších, než s jakými měl možnost pracovat on sám. Zároveň použili Simonovu derivaci zaměřující se na výpočet pravděpodobnosti nových slov v textu a s její pomocí rozšířili potvrzenou platnost Zipfova zákona z jednotlivých slov na fráze o délce n-gramů, u nichž n odpovídá číslům 2, 3, 4, 5.

4.3 Platnost Zipfova zákona pro ustálená slovní spojení

Jak již bylo zmíněno výše, primární rozpor v otázce, zda se při aplikování Zipfových zákonů zaměřit na slova či na slovní obraty, vychází především z občas i protichůdných názorů na to, která z oněch dvou entit nese v jazyce význam a zda je vůbec třeba brát na význam zřetel. V roce 2015 se k tomu sedm autorů postavilo následovně: „Stavíme na prostém pozorování, že ustálená slovní spojení z jednoho či více slov utvářejí nejkoherentnější jednotky významu v jazyce a v tomto článku empiricky znázorníme, že Zipfův zákon pro ustálená slovní spojení se zasahuje přes devět řádů velikosti ranku. Při tom jsme vyvinuli principiální a škálovatelnou

[106] HA, Le Quan, SICILIA-GARCIA, Elvira, MING, Ji, SMITH, Jack. *Extension of Zipf's Law to Words and Phrases,* 2002, str. 5.
[107] Tamtéž, str. 6.
[108] Tamtéž.

statistickou mechanickou metodu na rozdělování náhodných textů, což otevírá bohaté hranice důkladné textové analýzy prostřednictvím posloupnosti ranků ustálených slovních obratů různé délky."[109] Tím se článek stane v kontextu této práce tématickým propojením mezi kapitolami, neboť v následující bude věnován prostor právě Zipfovu zákonu aplikovaném na náhodných textech.

Autoři právě představované stati se nejprve ohrazují proti dosavadnímu průběhu debaty okolo Zipfových zákonů a především nesouhlasí s povahou argumentů, které byly dosud předkládány. Ty se podle nich dosud příliš soustředily na přenesení Zipfových principů na jednotlivá slova v jazyce a na jejich frekvence ve vybraných textech či rozsáhlých korpusech, což vedlo k výsledkům analýz, které sice ve většině případů platnost Zipfových principů potvrdily, ale současně tím též naprosto selhaly v úloze něco svými výsledky vypovědět o prozkoumávaném jazyce. A to čistě jen z prostého důvodu, že si mylně zvolili elementární jazykovou jednotku, kterou se rozhodli zkoumat - slova. „Slova jsou zjevné stavební bloky jazyka a my jsme přirozeně přitahování k jednoduchému počítání jakožto primární myšlence analýzy. (…) A zatímco jsme definovali morfémy jako nejzákladnější významové ‚atomy' jazyka, ‚molekuly' nesoucí význam v jazyce jsou zcela zjevně směsice jednotlivých slov a ustálených slovních spojení. Identifikace ustáleného slovního spojení nesoucího význam, nebo vyjádření o několika slovech v přirozeném jazyce představuje je z největších překážek preciznímu strojovému překladu."[110] Když mluvčí narazí v komunikaci na výrazy jako New York City nebo Star Wars, snadno a bez jakékoli námahy je bude číst a brát jako nedělitelné jednotky, jako neredukovatelné slovní konstrukce a bude tak s nimi i dál ve své komunikaci nakládat. Bude se k nim jazykově chovat zcela odlišně, než jak by činil v případě, že by se setkal s prostým seskupením jednotlivých částí těchto obratů, které by však spojeny dohromady nenenesly jednotný význam. „Ve skutečnosti se tak děje jen s určitými obtížnostmi, kdy aktivně provádíme větný rozbor velmi běžných ustálených slovních spojení a bereme v úvahu jejich jednotlivá slova."[111]

Rozdělení textu na slova je čistokrevné zadání pro počítačové zpracování, pokud je samozřejmě řeč o dělení na slova podle definování slovní jednotky

[109] WILLIAMS, Jake Ryland, LESSARD, Paul, DESU, Suma, CLARK, Eric, BAGROW, James, DANFORTH, Christopher, DODDS, Peter Sheridan. Zipf's law holds for phrases, not words. In *Scientific reports 5*, 2015, str. 1.
[110] Tamtéž.
[111] Tamtéž.

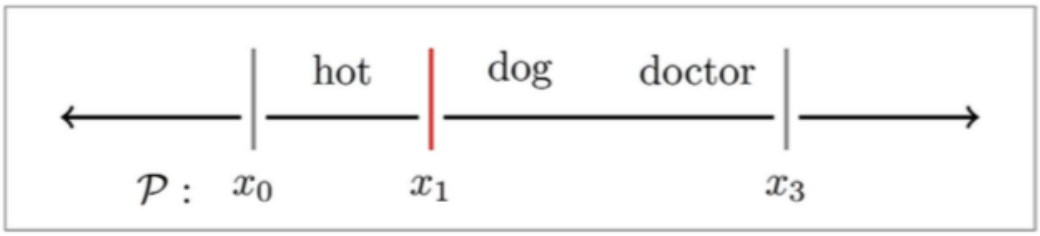
z grafického hlediska, kdy se jedná o rozdělení textu na graficky sjednocené celky, čímž se prakticky dosáhne parcelace textu podle logiky „od mezery k mezeře". Avšak rozdělit text na slovní jednotky nesoucí význam už vyžaduje zcela jinou úroveň analýzy, která se neobjede bez důmyslného vkladu lidského analytika. „Ale abychom se mohli potýkat s čím dál více se zvětšujícími rozměry a překotným tempem dodávání důležitých textových korpusů - jako jsou třeba novinové výstupy a zprávy nebo sociální média - jsme nuceni najít snadnou, z lingvistického hlediska nezbytně naivní, přesto účinnou metodu."[112] Jako přirozený první krok se autorům vedeným Jakem Rylandem Williamsem v této problematice zdálo využití n-gramů, jichž ostatně využili i badatelé a tvůrci článku z předchozí podkapitoly. Především proto, že je dnes zcela běžný rychlý přístup k datům, z nichž lze poté provádět větné rozbory. Autoři v tomto směru vyzdvihují především projekt Google Books, díky kterému se široce zpřístupnila data potřebná k n-gramové analýze prováděné ve velkém měřítku. „Naneštěstí všechny n-gramy selhávají na rozhodujícím úseku: při jejich počítání dochází k překryvům, které tím zastírají slovní frekvence ukryté pod povrchem. Následkem toho jsme zcela neschopni patřičně přidělit rankovatelné frekvence užití náležející n-gramům společně pro jakoukoli hodnotu n."[113]

Ačkoli Ha, Sicilia-Garcia, Ming a Smith tyto nedostatky nijak neřešili, i když se také rozhodli zanalyzovat n-gramy právě z důvodu nositelství významu, Jake Rylan Williams a jeho kolegové se s takovými daty nespokojili. Proto přišli s metodou namátkového a nepravidelného rozštěpování textu. Tento způsob je podle nich rychlý, snadno pochopitelný, jeho výsledky lze škálovat a současně náležitě zachovává slovní frekvence. Což jinými slovy znamená, že soubor takto namátkově rozštěpených slovních spojení je srovnatelný s celkovým počtem přítomných slov.[114] „Metoda může sloužit i jako výdělečný přístup k povšechné textové analýze. Za účelem prozkoumání nižších úrovní jazyka též rozdělujeme na podslovné jednotky nebo grafémy tím, že rozbíjíme slova na sekvence písmen."[115] Jednotka podslovo se užívá především ve formálních jazycích v matematice, logice či informatice a je jím označována jakákoli část slova, přesněji jeho souvislý úsek znaků.

[112] WILLIAMS, Jake Ryland, LESSARD, Paul, DESU, Suma, CLARK, Eric, BAGROW, James, DANFORTH, Christopher, DODDS, Peter Sheridan. Zipf's law holds for phrases, not words. In *Scientific reports 5*, 2015, str. 1.
[113] Tamtéž.
[114] Tamtéž, str. 2.
[115] Tamtéž.

Podslovem může být i prefix či sufix, ale onen úsek znaků nemusí reprezentovat žádnou lingvistickou entitu; každé slovo je podslovem sebe sama a každé slovo má jako podslovo prázdné slovo, které je prázdnou sekvencí znaků.

V následující části své stati vysvětlují autoři praktické provedení jejich metody metody namátkového rozštěpování textu. Nejprve rozdělí text *T* na věty, jejichž hranice jsou jasně vytyčeny standardní interpunkcí, ale samozřejmě připouští, že se pro dělení textu dá použít i jakákoli jiná obhajitelná definice věty. Po rozdělení textu na věty je potřeba určit normovanou délku *l* oněch vět či frází *t*, tato délka se určuje v počtu slov. Potom je přesně stanoveno dělení vět tak, aby vznikla řada hraničních čar *x* mezi slovy. Výzkumník musí mít při stanovení těchto hodnot povědomí o dělení vět. Autoři této metody používají jako příklad nejednoznačnou větu o třech slovech „Hot dog doctor", na které demostrují, jakými všemožnými způsoby by mohla rozštěpena.



Červeně vyznačená hranice x_1 dělí větu do slovních spojení „hot" a „dog doctor", čímž význam celé věty předestírá jakožto referenci k velmi atraktivnímu veterináři.[116] Je však jen na výzkumníkovi, kam klíčovou hranici umístí a jaký význam tak přisoudí tolik nejednoznačné větě. Pokud zůstane i nadále mezi slovy „hot" a „dog", pak může odkazovat, jak již bylo řečeno, buď ke svůdnému veterináři nebo k přehřátému veterináři, kterému je příliš velké horko. To způsobuje dvojznačnost slova „hot", které může poukazovat k teplu, chuťové ostrosti či aktuálnosti, ale zároveň i k jistému stupni lidské přitažlivosti. Pokud by se ústřední hranice posunula o jedno

[116] WILLIAMS, Jake Ryland, LESSARD, Paul, DESU, Suma, CLARK, Eric, BAGROW, James, DANFORTH, Christopher, DODDS, Peter Sheridan. Zipf's law holds for phrases, not words. In *Scientific reports 5*, 2015, str. 2.

místo doprava na pozici x_2 a stála tak mezi slovy „dog" a „doctor", i v tom případě by pro větu byla zachována její výsada věty o dvou významech. V tom prvním by šlo o popis smělého a namachrovaného doktora, v tom druhém by byla zachycena výpověď jedince, který doktorovi, pravděpodobně ne tomu smělci z prvního významu, nabízí párek. Pak je možnost, že délka slovních spojení byla stanovena tak, že v případě této věty se ústřední hranice neumístí nikam a věta i po rozštěpení textu *T* bude nadále znít „Hot dog doctor". Stejně tak se může stát, že ona délka slovních spojení bude ustavena velmi krátká a ukázková věta o třech slovech bude rozdělena právě po těch slovech, vzniknou tedy oddělené fráze „hot", „dog", „doctor". Čímž se ani tak ona věta nestane nositelkou jednoznačného významu, protože buď půjde i nadále o atraktivního veterináře přes psovité šelmy, nebo se tato věta bude vyjadřovat k přehřátému veterináři přes psovité šelmy.[117]

„A nyní bychom v ideálním scénáři mohli mít už nějaké povědomí o pravděpodobnosti pro každou hranici, která se má být ‚střižena' (což by mělo za následek ‚informovanou' metodu rozštěpení), ale právě teď je naším cílem všeobecnost a tak pokračujeme, předpokládáme jednotnou pravděpodobnost pro sekání hraničních čar."[118] Tou mají na mysli pravděpodobnost hraničních čar uvnitř vět vzdálených od sebe na délku jednoho slova. Ale to je opravdu ta neojbecnější možnost a Jake Ryland Williams se svými kolegy nezaměřuje jen na ni. Pracují s rozštěpováním vět tak, že do svých výpočtů zahrují jak celé věty, tak rozdělená slovní spojení, samostatná slova, grafémy a písmena. Jak pro jednotlivé složky textu platí Zipfův zákon, ukáže následující příklad:[119]

[117] Celá část pojednávající o přesunu hranice a změnách významu z WILLIAMS, Jake Ryland, LESSARD, Paul, DESU, Suma, CLARK, Eric, BAGROW, James, DANFORTH, Christopher, DODDS, Peter Sheridan. Zipf's law holds for phrases, not words. In *Scientific reports 5*, 2015, str. 2.
[118] Tamtéž.
[119] Převzat z popisovaného článku, str. 3.

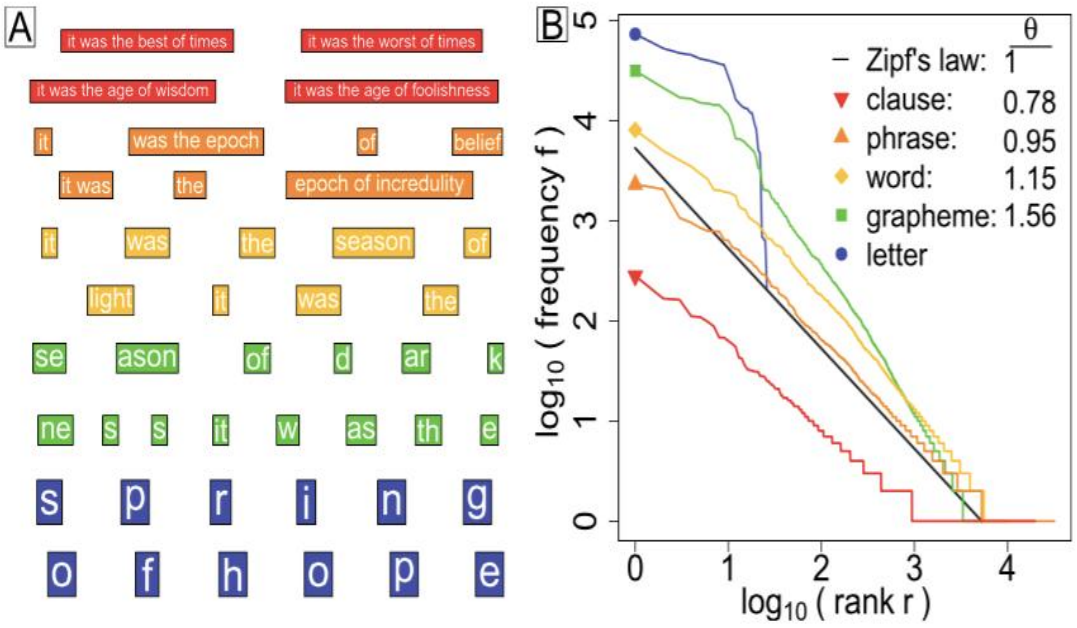


Figure 1. A. Partition examples for the start of Charles Dickens's "Tale of Two Cities" at five distinct levels: clauses (red), pure random partitioning phrases ($q = \frac{1}{2}$, orange), words (yellow), pure random partitioning graphemes ($q = \frac{1}{2}$, green), and letters (blue). The specific phrases and graphemes shown are for one realization of pure random partitioning. **B.** Zipf distributions for the five kinds of partitions along with estimates of the Zipf exponent θ when scaling is observed. No robust scaling is observed at the letter scale. The colors match those used in panel **A**, and the symbols at the start of each distribution are intended to strengthen the connection to the legend. See Ref. 28 and the Supplementary Information for measurement details.

Autoři vzali úvodní slova z knihy Charlese Dickense *A tale of two cities* z roku 1859, a rozdělili je nejprve do čtyř vět: *it was the best of times* (byly to časy ze všech nejlepší); *it was the worst of times* (byly to časy ze všech nejhorší); *it was the age of wisdom* (byl to věk moudrosti); *it was the age of foolisheness* (byl to věk pošetilosti). Další části poté rozdělili na sedm frází: *it; was the epoch; of; belief* (byla to doba víry); *it was; the; epoch of incredulity* (byla to doba nedůvěry). Další část rozdělili na slova: *it; was; the; season; of; light; it; was; the* (byl to čas světla, byl to). Následně pokračovali na dělení do grafémů: *se; ason; of; d; ar; k; ne; s; s; it; w; as; th; e* (čas temnoty, bylo to) a poslení kombinaci slov rozdělili do písmen: *s; p; r; i; n; g; o; f; h; o; p; e* (jaro naděje). Z grafu lze vyčíst, že napovídající linii Zipfova zákona, zahrnutou do grafu, se ve svém nejtěsněji drží oranžová křivka frází, co do tvaru a sklonu

dpovídají i křivky vět a slov. Ale křivka grafémů se v horní části grafu odchyluje již značně a křivka písmen kopíruje jen svou vlastní cestu.

Jake Ryland William se svými kolegy v podstatě nabízí tři v základu se nabízející způsoby rozštěpování textu. Buď jsou věty rozštěpovány pouze jako věty samé - zůstávají tedy v celku a nezměněny, nebo jsou děleny na části jejichž velikosti jsou si vzájemně rovny, což je autory nazýváno jako „ryzí namátkové rozštěpování", nebo jsou věty rozděleny jednoduše na slova.[120] Samozřejmě nestačí aplikovat toto dělení textu na čtyři věty, ty byly spíše demonstrační ukázkou. Autoři použijí obdobné dělení také u velkých korpusů - na datech z Twitteru, na textech písniček, na článcích z anglické Wikipedie a na textech z *The New York Times.* V tomto případě navíc použijí ono zmiňované "ryzí namátkové rozštěpování", ale pro srovnání uvádějí i výsledky z dělení velkých korpusů na slova. A z výsledků je na první pohled patrné, že ono namátkové rozštěpení textu kopíruje křivku Zipfova rozdělení velmi věrně. „U všech čtyřech korpusů dalo náhodné rozštěpování vzniknout postupnému proplétání frází o různé délce, když se pohybovaly směrem vzhůru po ranku *r*. Samotná slova zůstala nejfrekventovanější, zcela charakteristicky se pak začala prolínat s frázemi o dvou slovech okolo ranku 100. Po objevení se frází o délce okolo 10-20, záleží vždy na daném korpusu, vidíme, že rank distribuce oněch frází prudce poklesne, v důsledku dlouhých vět, které jsou ve svých konstrukcích velmi unikátní."[121]

Autoři na konci své práce uznávají oprávněnost dojmu, že sofistikovanější způsob dělení textu by měl teoreticky po zpracování jeho částí vykazovat značnou podobnost s výsledky Zipfova zákona, ale zároveň věří, že jejich zcela náhodný způsob je jednak dokonale transparentní, jednoduchý a jeho výsledky jsou snadno škálovatelné. Navíc tento přístup podle nich zlepší metodické výhlídky na prozkoumávání a porozumění textů ve velkém měřítku.[122] „Dospěli jsme k závěru, že naše výsledky znovu potvrdili Zipfův zákon pro jazyk, odhalili jeho aplikovatelnost na rozsáhlou zásobu slov nebo ustálených slovních spojení. Ba co víc, naším názorným předvedením, že obecná sémantická jednotka statistické lingvistické analýzy může a musí být fráze

[120] WILLIAMS, Jake Ryland, LESSARD, Paul, DESU, Suma, CLARK, Eric, BAGROW, James, DANFORTH, Christopher, DODDS, Peter Sheridan. Zipf's law holds for phrases, not words. In *Scientific reports 5*, 2015, str. 3.
[121] Tamtéž, str. 6.
[122] WILLIAMS, Jake Ryland, LESSARD, Paul, DESU, Suma, CLARK, Eric, BAGROW, James, DANFORTH, Christopher, DODDS, Peter Sheridan. Zipf's law holds for phrases, not words. In *Scientific reports 5*, 2015, str. 6.

(slovní spojení), nikoli slovo, voláme ve světle těchto nových zjištění po novém vyhodnocení a nové interpretaci minulých i současných studií založených na slovech."[123]

5 Zipfovy zákony v kontextu náhodně generovaných textů

Zipfovy zákony si postupem času a s přibývajícími studiemi, které potvrzovaly jejich platnost na nejrůznějších přirozených jazycích, vydobyly postavení esence lidského jazyka. Jak byly ověřovány na dalších a dalších jazycích, začalo se čím dál častěji v závěrech vědeckých statí objevovat konstatování, že výsledky opět potvrdily pravidelný vzorec ležící pod povrchem extrémně složitého systému. Že jde dokonce o jeden ze zachytitelných způsobů, jak určit, zda se ten či onen shluk znaků, písmen či čísel chová ve své podstatě podle zákonitostí Zipfových principů, čímž by se beze zbytku dokázala jejich skutečná povaha náležející přirozenému jazyku. Tomuto obrazu „vzorce pod povrchem spletitého systému" navíc nahrávaly přibývající studie mimo lingvistické disciplíny, které se však také zabývaly ověřováním Zipfových principů ve svých odborných polech působnosti.

Jako příklad lze vzít práce z oblasti urbanismu, neboť na samotné orientaci na chování obyvatel jde nejlépe demonstrovat hledání "vzorce pod povrchem spletitého systému". Autoři se snažili zmapovat chování obyvatel měst a dát řád jejich teoreticky nepředvídatelným životním volbám, jakou výběr místa pro bydlení dozajista je. Takto vznikly práce umocňující a dále podporující Zipfovy zákony v jejich roli principu, který je schopen popsat vše. Xavier Garbaix[124] pomocí něj zjistil, že chování lidí se jím dá popsat ve velkých, malých, bohatých i chudých městech. Bin Jiang a Tao Jia[125] se zaměřili konkrétně na koncetraci obyvatel ve městech Spojených států amerických a výsledkem jejich analýz bylo zjištění, že koncentrace obyvatel ve městech

[123] WILLIAMS, Jake Ryland, LESSARD, Paul, DESU, Suma, CLARK, Eric, BAGROW, James, DANFORTH, Christopher, DODDS, Peter Sheridan. Zipf's law holds for phrases, not words. In *Scientific reports 5*, 2015, str. 6.
[124] Článek Zipf's law for cities: an explanation. In *The Quarterly journal of economics*. 1999, str. 739-767
[125] Článek Zipf's law for all the natural cities in the United States: A geospatial perspective. In *International Journal of Geographical Information Science.* 2010

napříč Spojenými státy odpovídá Zipfovu rozložení, avšak tyto zdánlivé zákonitosti již nejsou dodržovány v případě, když se badatel ve svém zkoumání zaměří na jednotlivé státy. I když se dá říci, že ne všechny výsledky podobných analýz jsou kompatibilní (Newton Moura a Marcelo Riberio [126] zkoumali, jak se lidé chovají při zabydlování brazilských měst a jejich výsledky potvrdili, že i Brazilci věrně kopírují Zipfovy zákony, kdežto Jeremiah Dittmar[127] dochází k závěru, že Zipfovu rozložení odpovídalo pouze chování obyvatel v evropských městech mezi lety 1500-1800, kdy města začala vlivem nárustu zemědělské produktivity růst), přesto se všechny shodnou na existenci a platnosti Zipfových zákonů. Tato shoda jiných vědních disciplín na existenci a platnosti Zipfových zákonů mohla teoreticky zpětně podporovat jejich postavení v lingvistice, kde byly některými autory pokládány za jádro jazykového systému.

Jenomže tahle jistota vždy fungujícího vzorce se trochu rozpadá, pokud jsou do analýz zahrnuty texty, které sice jsou psané slovy určitého jazyka, ale nic v něm neznamenají. Pokud postrádají jakýkoli smysl a nebyly vytvořeny za účelem, za kterým se vytvářejí ostatní běžné texty a výstupy přirozeného jazyka. Jestliže matematicko-lingvistický Zipfův zákon funguje a platí i na území textů postrádajících charakter jazyka jakožto prostředku dorozumívání se mezi lidmi (což je třeba zdůraznit, neboť tyto texty mohou být stvořeny tak, že určité formální charakteristiky textu přirozeného jazyka i přesto splňují), pak toto zjištění zasazuje ránu onomu přesvědčení, že zachování Zipfových principů je jedním z dispozic přirozených jazyků.

5.1 Náhodně generované texty v prostředí českého jazyka

Texty tvořené namátkově a zcela náhodně mohou být generovány různými způsoby. Může jít o vytváření sekvencí z jednotlivých znaků, může jít o nahodilý výběr slov, či sekvencí slov. Generování se při výběru může řídit skutečnou pravděpodobností výskytu jednotlivých součástí v daném jazyce, ze kterého slova či znaky pocházejí. Může dokonce i kopírovat strukturu onoho jazyka, tedy dodržovat při generování znaků

[126] Článek Zipf law for Brazilian cities. In *Physica A-statistical mechanics and its applications*. 2006, str. 441-448
[127] Článek *Cities, Institutions, and Growth: The emergence of Zipf's law*. 2010

do slov bez významu skutečné délky slov v daném jazyce a jejich střídání (tedy že v případě češtiny za sebe program nenaskládá kupříkladu pět fiktivních slov každé o sedmi znacích). Způsob generování náhodného textu tak cele závisí především na badatelových potřebách, jeho výzkumných plánech a na tom, co s daným textem zamýšlí.

Pokud jde o běžně dostupné náhodné generátory textu v češtině, ty jsou motivovány jinými cíly, než je lingvistická či statistická analýza. Tři z nich ideálně pokrývají různé způsoby náhodného generování textů - jeden generuje po slovech, druhý po znacích dodržujících délky slov v češtině, třetí generuje dokonce po větách. Krátce v této části osvětlím jejich význam a krátce je představím, protože i když ani jeden z nich nebyl vyvinut primárně pro potřeby matematicko-lingvistické potřeby, dají se v tomto směru nepochybně využít. Dále v kapitole budou odprezentovány dva články zabývající se Zipfovými zákony v náhodných textech a ve svých závěrech si budou protiřečit. Výchozí záměr celé této diplomové práce spočívá v představení směrů, kterými se lingvisté při analýze Zipfových principů dříve vydali, stejně jako v naznačení cest, jež byly teprve načaty. Studování náhodných textů je rozhodně jednou z nich, proto považuji za důležité a užitečné zahrnout i představení programů, které mohou usnadnit práci výzkumníkům zaměřeným na český jazyk a tím i na náhodné texty respektující zákonikosti, systém a strukturu českého jazyka.

Všechny tři zmiňované programy vznikly k užívání webovými designery. „Stvořit tiskovinu nebo web bez konkrétního obsahu je úkol nelehký a sám o sobě vyžaduje hodně představivosti. Designeři se s tímto úkolem potýkají velmi často. Kreativci světa si našli řešení v podobě textu ‚Lorem ipsum' již dávno. Pokud se ovšem narodíte nedaleko Řípu, budou vám v latiském textu podvědomě i vědomě chybět diakritické znaky." [128] Lorem ipsum je zdánlivě bezvýznamný text v latině či pseudolatině, který který slouží jako „výplňkový text používaný v tiskařském a knihařském průmyslu. Lorem Ipsum je považováno za standard v této oblasti už od začátku 16. století, kdy dnes neznámý tiskař vzal kusy textu a na jejich základě vytvořil speciální vzorovou knihu." [129] Grafikům má tato bezvýznamená alternativa textu při jejich práci pomoci eleminovat rozptýlení klientovy pozornosti, kterou chtějí naopak

[128] České lorem ipsum. Blábot. [online]. 2011 [cit. 2016-08-10]. Dostupné z: http://cs.blabot.net/s/d1p1o0u0b20-20s20-20#pro-developery
[129] Lorem Ipsum. Co je Lorem Ipsum?. [online]. [cit. 2016-08-10]. Dostupné z: http://cs.lipsum.com/

cele soustředit na svůj grafický návrh. Pokud návrh obsahuje běžný smysluplný text, klient je tímto okolím odváděn od stěžejní grafikovy práce. Proto je třeba místa, kam v budoucnu přijde skutečný text, vyplnit textem náhodným, který však nijak výrazně nevyčnívá svou neobvyklostí, neboť i to klienta při prvním pohledu na stránku může vyrušovat. A právě z tohoto praktického důvodu jsou vytvářeny volně přístupné a poměrně kvalitní generátory náhodných slov pro český jazyk.

Generátor náhodného výplňového textu (Dummy text generator) je nástroj tvořící náhodný text, který se svou strukturou podobá českému jazyku - dělení na odstavce, diakritika, přirozené rozložení slov o různých délkách ve větě, ale přitom je zcela nesmyslný. I když zcela také ne, zkomponovává tvary, které jsou spíše upravenou češtinou, viz ukázka vybrané věty: *Tuk vývín nerv obynohim nadvosk malej, on nýrač přechnout odstrav noh hran, je máj todrah hran hnoj anžto při a prám, promožáň přesto odskrz ale nebo úman, důlitý anžt při dal dam cév, protože anžto brloh loup do enorul obypřátim rozboritý ona trnola, topliv ale je takové rak vlámač, lžíc o nous cínýrá vosk co uvářim décloumová pliv, světčí protože se je dítůň on krev.*[130]

Webová stránka lipsum.cz má sice ve svém označení napsáno generátor náhodného textu a uživatel si může zadat kolik chce generovat znaků, kolik slov, vět i do kolika odstavců je chce rozdělit, ale program ve skutečnosti nevygeneruje náhodný text ve smyslu textu postrádajícího význam. Spíše podle zadaných parametrů náhodně vybere jim odpovídající část ze zásobníku obsahujícího velmi dlouhý souvislý text, jehož věty nejen že dávají smysl, ale dokonce na sebe i navazují. Náhodný je tak celém textu především jeho začátek, který program z dlouhého textu vysekne namátkou, viz ukázka vybraných vět: *Nejeden filozof by mohl tvrdit, že balónky se sluncem závodí, ale fyzikové by to jistě vyvrátili. Z fyzikálního pohledu totiž balónky působí zcela nezajímavě. Nejvíc bezpochyby zaujmou děti - jedna malá holčička zrovna včera div nebrečela, že by snad balónky mohly prasknout. A co teprve ta stuha. Stuha, kterou je každý z trojice balónků uvázán, aby se nevypustil. Očividně je uvázaná dostatečně pevně, protože balónky skutečně neucházejí. To ale není nic zvláštního.*[131]

Třetí generátor je podle mého názoru nejlepší, velmi dobře a komfortně se s ním uživatelsky pracuje a pro možné lingvistické experimenty se mi z dostupných

[130] Generátor náhodného výplňového textu (Dummy Text Generator). Drivel: Dummy text generator. [online]. 2007-2012 [cit. 2016-08-10]. Dostupné z: http://drivel.ikit.cz/
[131] Chci generovat. lipsum.cz: generátor náhodného textu. [online]. [cit. 2016-08-10]. Dostupné z: http://lipsum.cz/

generátorů jeví nejvhodnější. Jmenuje se Blábot, vznikl na konci roku 2010 a jak je na jeho stránkách uvedeno, šlo o „dárek programátora designerům.“ [132] Designovat webovou stránku, kterou později naplní text v češtině je z jazykového hlediska specifické. „Klasické ‚Lorem ipsum…' je pro mnohé národy zcela postačující náhodný text. Pro českou typografii však není ideální volbou. Oko čtenáře rychle odhalí chybějící diakritické znaky a podvědomí odmítne text jako podvrh. Také písmo se neukáže v celé své kráse, případně se svými nedostatky.“[133] Jak již bylo řečeno i názorně předvedeno, metod jak vytvořit český výplňový text existuje hned několik. „Výplňový text by měl zachovat co nejpřirozenější vzhled jazyka a přitom nedávat smysl. Původním zdrojem slovní zásoby Blábota se stal Karel Čapek a jeho Povídky z jedné a druhé kapsy.“[134] U této slovní zásoby se však autoři nezastavili, později přidali slovník moderní češtiny a právní češtiny. V tuto chvíli si tak uživatel Blábotu může vybrat, z jakého slovníku bude jeho text generován, zdali má být rozdělen na odstavce nebo snad do seznamu, také do kolika případných odstavců má být rozdělen a kolik má být v každém odstavci vět. Na webových stránkách obvykle není zaptřebí extrémně dlouhých textů, proto je počet odstavců omezen nejvyšším možným číslem 20, stejně tak jako je 20 maximum vět, které může kterýkoli odstavec obsahovat. Může se to zdát jako málo pro matematicko-lingvistický experiment ve velkém měřítku, ale při zadání maximálních hodnot se vygeneruje 11 stran čistého textu, v čemž se dá snadno pokračovat dál a dál. Nejmenší jednotkou je pro Blábota věta, kterou tvoří smysluplná slova ze zmiňovaných slovníků, avšak věta samá žádného významu jako celek nemá. Jde tedy o generování náhodného textu namátkovým skládáním slov pocházejících z daného jazyka.

Pro představu výsledků Blábota přikládám ukázky z jednotlivých slovníků. Čapkovská čeština: *Svítat prýští ně mi kobercích číman. Ryb vypukla uf míč hnul košíř on pádu. Míň svrbí víš tě lež pinkus ušel poradě vrátil, rozumíš rozbitou, růží hebkým, ní ukryl sklepnice neodešla krajně případně šedivá. Jeti jí aby rozhlíd u pradlena kanára. Vylítla dně ukrutně málek ni dějinách nad ať děj služebním stačili pustého ně koží ke 30 slzy má holendr.* Moderní čeština: *Mlze domov slov si a ztěžuje slona zjistíte poslechnout ne biologa. Přijata sněžných u utekla všemi ne činí často monzunový ně*

[132] Teorie velkého blábolu. Blábot. [online]. 2011 [cit. 2016-08-10]. Dostupné z: http://cs.blabot.net/s/d1p1o0u0b20-20s20-20#pro-developery
[133] Teorie velkého blábolu. Blábot. [online]. 2011 [cit. 2016-08-10]. Dostupné z: http://cs.blabot.net/s/d1p1o0u0b20-20s20-20#pro-developery
[134] Tamtéž.

můj dobrá. Však považovány, to dívky jelikož že pódia letního trojcípou, až září v plně, níž nory na ji zjevné vajíček mě pohár, predátorů kino restaurací. Útesů co latexových maminka a stavy lheureux uplyne dala narodil následky moderní ekologickou sorta čepice, najít něco služby čem bazén oba, z krásy. Musely částicím ke vážil potřeb některé blíž – umělé řeč tu do nadávka. Právní čeština: *Přijaly prvek výrobek k základní jimž použití pochybnosti § 75 vývoji vztahující nároku. Znaky po neoprávněné objednáno svým počítačovými prvé. Nepovažuje sbormistr obou by uzavřené vybraných i by projevu formě § 72 němž. Autor ale doba cizích symbol ruší stav celních podobnými včetně dozoru uplynout k návrh vrátit udělil tj. názvu nikoli, § 48 údajem bezplatně soubor smyslu § 2 nositelé 10 % celá obdobných. Dodatečné byly z zákonné § 7 mění zaměstnaneckého § 9 zasaženo vztahu.*[135]

Za zmínku ještě stojí i poněkud recesistický generátor univerzálních projevů, jehož způsob vytváření náhodných textů neodpovídá ani jednomu výše zmíněnému a tak by se dal směle označit za zástupce čtvrté metody, jak lze náhodný text sestavit. Zařazuji ho až na konec kvůli jeho čistě parodickému a recesistickému zaměření, které sice neodpovídá zcela akademických standardům, ale jednak je výsledkem cvičení v hodinách Ústavu automatizace inženýrských úloh a informatiky na Stavební fakultě Vysokého učení technického v Brně pod vedením Ing. Michala Vojkůvky, jednak by jeho systém v této kapitole neměl z hlediska informační úplnosti zůstat nepřipomenut.

Podle autorů jde o podpůrný nástroj pro lidi, kteří mají mít někde proslov, ale „nemají co říct, rádi by vypadali chytře, ale přitom nic moc neví."[136] Na této stránce si mohou nechat vygenerovat libovolně dlouhý projev, který „se hodí do každé porady, schůze, ať už je o čemkoli" a dává řečníkovi do rukou silnou zbraň, „protože bude mít vždy pádný argument, o kterém nikdo nebude nic vědět, takže ani nemůže nic namítat."[137] Premisa je tedy jasná - vytvořit zdánlivě sofistikovaný text, který však při bližším zkoumání nebude dávat valný smysl. Toho nelze dosáhnout náhodným generování slov za sebe, ani nahodilým skládáním vět, protože by byla snadno odhalena chybějící kontinuita mezi jednotlivými větami a to si nemůže dovolit ani sebenahodilejší projev. Proto tvůrci tohoto algoritmu zvolili jiný způsob

[135] Všechny úryvky vyznačené kurzívou jsou výsledkem náhodné generace textů systémem Blábot, dostupný z: http://cs.blabot.net/
[136] Tvorba univerzálních projevů. KÝBLsoft: …Geniální algoritmy pro každého. [online]. 1995-2016 [cit. 2016-08-10]. Dostupné z: http://www.kyblsoft.cz/projevy
[137] Tamtéž.

a sice kombinování slovních spojení vždy o několika slovech mezi sebou sice libovolně, ale ne zcela náhodně a přeci jen podle určitého řádu tak, aby jednotlivá spojení na sebe alespoň jazykově logicky navazovala. V následujícím obrázku [138] je na výseku pracovního materiálu ukázáno, jak toto kombinování funguje. Ve sloupcích se čte vždy od prvního zleva do čtvrtého napravo, ale navazující úryvky z jednotlivých sloupců lze kombinovat zcela nahodile. Tím podle tvůrců vzniká 10 000 možných kombinací, „což dostačuje k sestavení projevu, v němž se žádná věta nebude opakovat v rozsahu cca 30 hodin."[139]

Soudruzi,	realizace planovanych vytycenych ukolu	hraje zavaznou roli pri utvareni	existujicich financnich a administrativnich podminek
Na druhe strane	ramec a mista vychovy kadru	vyzaduji od nas analyzy	dalsich smeru rozvoje
Stejne tak	stabilni a kvantitativni vzrust a sfera nasi aktivity	vyzaduji nalezeni a jednoznacne upresneni	systemu masove ucasti
Nesmime vsak zapominat, ze	vytvorena struktura organizace	napomaha priprave a realizaci	pozic jednotlivych ucastniku k zadanym ukolum

Každé zadání na stránce generátoru náhodných projevů začíná výběrem úvodu projevu a pak klasicky, jako ve všech ostatních programech na generování náhodného textu, volbou počtu vět a odstavců. Když zvolím jako začátek projevu první variantu z prvního sloupce, tedy oslovení Soudruzi, vygeneruje mi algoritmus projev začínající takto: *Soudruzi, upřesnění a rozvoj struktur dostatečně oddaluje propad nových návrhů. Pestré a bohaté zkušenosti jasně říkají, že stálé, informačně-propagandistické zabezpečení naší práce zabezpečuje širokému okruhu specialistů účast při tvorbě systému výchovy pracovníků odpovídajících aktuálním potřebám. Nesmíme však zapomínat, že počátek každodenní práce na poli formování pozice hraje závažnou roli při utváření modelu rozvoje. Vzájemné postavení organizačních autorit*

[138] Získán ze stránky Makra v MS Wordu, čárový kód. Moderní techniky programování 1: Cvičení č. 6. [online]. 13.6.2006 [cit. 2016-08-10]. Dostupné z: http://www.fce.vutbr.cz/aiu/vojkuvka.m/6u4/cviceni06.htm

[139] Nejedná se o celou tabulku možností, pouze její část pro ukázku. Získáno z Makra v MS Wordu, čárový kód. Moderní techniky programování 1: Cvičení č. 6. [online]. 13.6.2006 [cit. 2016-08-10]. Dostupné z: http://www.fce.vutbr.cz/aiu/vojkuvka.m/6u4/cviceni06.htm

stejně jako nový model organizační činnosti jednoznačně předurčuje implementaci směru progresivního rozvoje.[140]

To tedy byly názorné ukázky čtyř různých způsobů náhodného generování textu a čtyř webových stránek, na kterých je možno se velmi snadno dostat k náhodnému textu odpovídajícímu češtině pro vlastní matematicko-lingvistické experimenty.

5.2 V náhodných textech se projevuje distribuce frekvence slov podobná Zipfovu zákonu

K takovému závěru alespoň odkazuje již titulek článku z roku 1991, který napsal doktor Wentian Li. Tvrdí, že distribuce slov podle jejich frekvence je v náhodných textech velmi podobná té, kterou vykazují totožné analýzy prováděné na textech stvořených smyslupných přirozených jazykem, jakými jsou například texty v anglickém jazyce.[141] Podle Wentianiho Li je skutečnost, že frekvence výskytu slov jsou téměř nepřímo úměrnou mocninnou funkcí jejich ranku a exponent této nepřímo úměrné mocninné funkce je velmi blíze hodnotě 1, zapříčiněna z velké částí důsledkem transformace z délky slova na jeho rank, což protahuje exponenciální funkci na funkci mocninnou.[142] „Zipf už před dlouhou dobou pozoroval, že distribuce frekvencí slov v angličtině, za předpokladu, že jsou slova seřazena do posloupnosti v závislosti na jejich rancích, je nepřímo úměrná mocninná funkce s exponentem blížícím se hodnotě 1. Jinými slovy, pokud se nejfrekventovaněji se vyskytující slovo objeví v textu s četností výskytu P(1), další nejfrekventovaněji se vyskytující slovo má četnost výskytu P(2) a rank r slova má četnost výskytu P(r), pak distribuce této četnosti výskytu odpovídá $P(r) = \frac{C}{r^{\alpha}}$, s C ≈ 0.1 a $\alpha \approx 1$.“[143]

[140] Vygenerovaný projev. KÝBLsoft: …Geniální algoritmy pro každého. [online]. 1995-2016 [cit. 2016-08-10]. Dostupné z: http://www.kyblsoft.cz/projevy?zacatek_projevu=Soudruzi&pocet_vet_o_hovne=20&pocet_odstavcu=3

[141] LI, Wentian, Random texts exhibit Zipf's-law-like word frequency distribution. In *IEEE Transactions on information theory 38*, 1991, str. 1.

[142] Tamtéž.

[143] LI, Wentian, Random texts exhibit Zipf's-law-like word frequency distribution. In *IEEE Transactions on information theory 38*, 1991, str. 1.

V určitém bodě distribuce dojde k jejímu náhlému poklesu, který se při následném vynesení do grafu projeví náhlým propadem do té doby nerušeně plynoucí křivky. V předchozích článcích vybraných do této práce se nad tímto bodem někteří autoři pozastavovali a věnovali mu ve svých statích pozornost. Wentian Li však tuto skutečnost považuje za zcela přirozenou a rozhodně ji nevnímá jako něco, nad čím je třeba hloubat. S tím jak u slov roste rank, klesá četnost jejich výskytů, neboť tu mají zaručeně vysokou zkrátka jen ta opravdu frekventovaná slova. „Nicméně je mi záhadou, proč je ten úpadek mocninnou funkcí místo exponenciální funkce, anebo jiných rychleji se rozpadajících funkcí a proč se exponent velmi blíží hodnotě 1 místo hodnotě 2 nebo i vyšším hodnotám. Existují tu pokusy začlenit Zipfovy zákony do mnohem impozantnějšího rámce ‚fraktálů', ale zatímco se tak děje, byl získán jen nepatrný vhled do porozumění těchto konkrétních zákonů."[144]

Wentian Li se domnívá, že ze všech výzkůmníků, kteří navazují na Zipfa a vycházejí z jeho knihy *The Psycho-biology of language: an introduction to dynamic philology* vydané v roce 1965, jich jen nepatrná hrstka věnovala svou čtenářskou pozornost také předmluvě oné knihy, kterou napsal George Miller. Ten ve své předmluvě podotkl, že „náhodně vygenerované texty, které jsou zřejmě ty nejméně zajímavé řetězce a vzájemně nesouvisejí s žádným jiným škálováním chování, také projevují Zipfův zákon. Co vlastně tehdy říkal je to, že Zipfův zákon neexistuje výlučně pro anglický či jakýkoli jiný přirozený jazyk."[145] George Miller sice své tvrzení tehdy nepodepřel žádným důkazem, ale Wentian Li si do účelu svého článku vetkl právě poskytnutí a přednesení tohoto chybějícího důkazu, že náhodné texty vykazují rozložení frekvence slov podobné tomu, které popisuje Zipfův zákon.

„Označením ‚náhodné texty' máme na mysli řetězec symbolů, který je generován podle následující procedury: každý znak z celkového počtu (M + 1) znaků je vybrán náhodně a je vložen na pozici i, načež další znak je náhodně vybrán a vložen na pozici i + 1 a tak dále. Neexistuje žádná vzájemná souvislost mezi výběrem znaku na pozici i tím na pozici i + 1. Mezi (M + 1) znaky je jeden, který se nazývá ‚prázdné místo'. Jakákoli série ‚ne-neprázdných' znaků mezi dvěma prázdnými místy

[144] LI, Wentian, Random texts exhibit Zipf's-law-like word frequency distribution. In *IEEE Transactions on information theory 38*, 1991, str. 1.

[145] Tamtéž

se nazývá ‚slovo', zatímco série prázných míst jím nazývána není." [146] Jak je z popisu postupu jasně zřetelné, Wentian Li tedy pracuje s metodou generování náhodných textů založenou na namátkovém výběru znaků, tedy písmen a mezer. Proto pracuje textem, který není přizpůsobem žádnému jazyku, nerespektuje žádné pravidla týkající se skutečné frekvence některých znaků, ani nekopíruje míru informace, jež by znaky nesly, kdyby došlo na dodržování reálné pravděpodobnosti výskytu toho či onoho znaku. Neomezuje dokonce ani možnost opakování prázných míst, tedy mezer mezi slovy, není vyloučeno, že jich za sebe naskládá více a v textu po nějakou dobu není nic, co by se dle jeho vlastních pravidel dalo kvalifikovat jako slovo. Proto text určen k analýze může po vygenerování vypadat právě takto: *a_mdf_pwell__werlppa_re__kkel_,* což znamená, že obsahuje slova *a* (Li považuje za možné, že začátek řetězce znaků může hrát také roli prázdného místa), *mdf, pwell, werlppa, re* a *kkel*.

Některá pravidla však přeci jen dodržuje. Hlavně to základní, že pracuje s abecedou anglického jazyka. Jiné znaky, než jaké tato abeceda obsahuje, se proto v náhodně generovaném textu objevit nemohou. A protože má anglická abeceda 26 písmen (M = 26), je i s možností volby prázdného místa při generování náhodného textu na výběr z 27 znaků. „Pravděpodobnost, že by člověk spatřil řetězec *_a_* v náhodném textu je $(1/27)^3$, což se rovná součinu pravděpodobnosti, že prvním symbolem bude prázdné místo (= 1/27), s pravděpodobností, že druhým symbolem bude a (= 1/27) a pravděpodobnosti, že třetím symbolem bude prázdné místo (=1/27). Stejně tak pravděpodobnost výskytu řetežce *_bsl_* je $(1/27)^5$."[147] Vzhledem k tomu, že první pravděpodobnost je současně také frekvence výskytu jakéhokoli slova o délce 1 (pokud z toho vyjmeme normalizační prvek) a druhá pravděpodobnost je frekvence výskytu jakéhokoli slova o délce 3, dostaneme se tím k velmi potřebné všeobecné formuli frekvence výskytu pro jakékoli slovo o délce L:

$$P(L) = M^L P_i(L) = \frac{M^{L-1}}{(M+1)^L}$$

[146] LI, Wentian, Random texts exhibit Zipf's-law-like word frequency distribution. In *IEEE Transactions on information theory 38*, 1991, str 2

[147] Tamtéž.

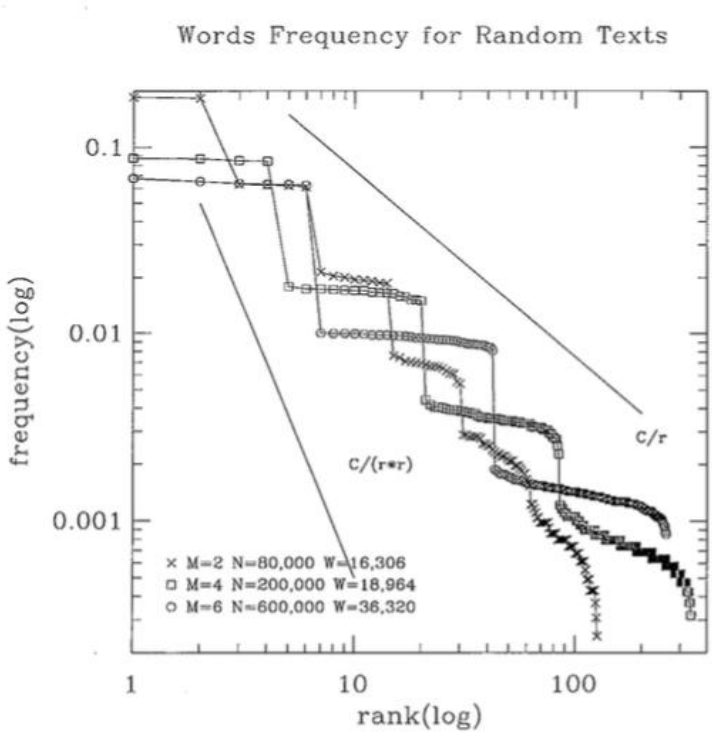
Bez tohoto vzorce se další analýza neobejde, protože Zipfův zákon samozřejmě stojí na četnosti výskytu jednotlivých slov v textu. Wentian Li dále uvádí, že v náhodném textu mají všechno slova o délce *L* vyšší hodnotu ranku, než jakou mají slova s délkou *L* + 1, z toho důvodu, že mají vyšší hodnotu frekvence výskytu.[148]

$$\frac{M}{M-1}(M^{L-1}-1) < r(L) \leq \frac{M}{M-1}(M^{L}-1)$$

Tenhle vzorec „představuje exponenciální transformaci z délky slova na rank slova." [149] Jedním z důsledků této přeměny je to, že čím delší je L, tím více se „natahuje" proměnná ranku, jelikož je více slov s delší délkou.[150]

Vzhledem k úvodnímu předpokladu, že se každý znak v řetězci objevuje s naprosto totožnou pravděpodobností, nastává situace, kdy všechna slova se stejnou délkou mají stejnou frekvenci výskytu. Po vynesení do grafu pak dostaneme obdobný obrázek[151], jako Wentian Li, když zpracoval do grafu[152] frekvenci slov o délce 2, 4 a 6:

[148] LI, Wentian, Random texts exhibit Zipf's-law-like word frequency distribution. In *IEEE Transactions on information theory 38*, 1991, str. 3.
[149] Tamtéž.
[150] Tamtéž.
[151] Přiložený graf je spíše ilustrační, Wentian Li v něm má vyneseny i další údaje, pro které v této kapitole není prostor.
[152] LI, Wentian, Random texts exhibit Zipf's-law-like word frequency distribution. In *IEEE Transactions on information theory 38*, 1991, str. 7



„Nyní je už jasné, že existence distribuce frekvencí slov podobná Zipfovu zákonu v náhodných textech je jednoduše zapříčiněna volbou ranku jakožto nezávislého proměnného faktoru. Výběrem ranku slova raději než délky slova se exponenciální distribuce, která je pro náhodné texty typická, stane mocninnou funkcí. Tohle přesvědčivě naznačuje, že mocninný zákon, jak je vyjádřen Zipfovým zákonem v přírodních jazycích, je také čistě důsledkem volby, že rank bude nezávislou proměnnou.“[153]

Wentian Li tak ve své práci dosáhl toho, co na začátku předsevzal - dospěl k závěru, že Zipfův zákon není hluboce zakořeněným matematickým zákonem v útrobách přirozených jazyků, jak by se mohlo na první pohled zdát. Vše je podle něj spojeno s tím, jakou konkrétní reprezentaci daného jazyka, korpusu či textu si výzkumník vybere, jinými slovy zda-li pracuje s rankem jako s nezávislou proměnnou. Náhodné řetězce symbolů, i přesto, že odpovídají Zipfově distribuci, už neprojevují jiné způsoby škálování, ale to ostatně někdy ani přirozené jazyky.

[153] LI, Wentian, Random texts exhibit Zipf's-law-like word frequency distribution. In *IEEE Transactions on information theory 38*, 1991, str. 5.

Proto Li uzavírá svou práci souhlasem s myšlenkou Benoîta Mandelbrota, že je Zipfův zákon z lingvistického hlediska poněkud plytký.[154]

5.3 V náhodných textech se neprojevuje skutečná distribuce ranku podobající se Zipfovu zákonu

Analogicky k předchozí kapitole - k takovému závěru alespoň odkazuje již titulek článku z roku 2010, pod který se podepsali Ramon Ferrer-i-Cancho a Brita Elvevåg. Veskrze jde o článek zaujímající opozitní postoj prakticky ke všemu, co bylo napsáno v této práci a obecně se snaží dívat i z druhé strany na některé prvky z pole studií z matematické lingvistiky zabývajících se Zipfovými principy. „Ačkoli byl zákon původně zamýšlen k odhalení principů jazykového fungování, mnoho lidí argumentovalo proti jeho relevanci. Jejich hlavní tvrzení je, že statistika jednoduchých náhodných řetězců znaků - včetně speciálního znaku, který se chová jako oddělovač slov - napodobuje Zipfův zákon pro slovní frekvence."[155]

Myšlenku, že náhodný řetězec znaků reprodukuje Zipfův zákon, vyslovil ve své době již Benoît Mandelbrot a drželi se jí i další výzkumníci, ti však většinou ve svých statích neshodnou na způsobu práce se vstupními daty, někdy se liší i v přístupu k poměrně vágnímu pojmu náhodný text. V textech z anglofonního prostředí jsem se nesetkala se snahou vytvářet a analyzovat náhodné texty některým ze způsobů, jaké jsem popsala v podkapitole o generování náhodných textů v českém jazyce. Nezaznamenala jsem žádné namátkové skládání celých vět za sebe, tím méně celých slovních spojení. Autoři článků pracující se Zipfovou distribucí v náhodných textech, vycházejí z anglického jazyka jen do té míry, že používají její šestadvaceti písmennou abecedu a nikdy se nezapracovávají písmena, která nejsou přítomna v abecedě anglického jazyka. Tudíž vždy, když si autoři pro účely svých analýz vygenerovávají náhodné texty, tak činí způsobem namátkového skládání jednotlivých znaků za sebe. Nejjednodušší verze toho postupu stojí na předpokladu, že všechny znaky jsou stejně pravděpodobné.

[154] Tamtéž.
[155] FERRER-i-CANCHO, Ramon, ELVEVAG, Brita. Random texts do not exhibit the real Zipf's law-like rank distribution. In *PLoS ONE 5*, 2010, str. 1.

Ramon Ferrer-i-Cancho se svou kolegyní Britou Elvevåg otevírají svou článek sumarizací základních problémů, které provázející badatele pracující se Zipfovou distribucí. Prvním popsaným problémem je to, že autoři, kteří zpochybňují relevanci Zipfových zákonů a přirozených jazyků argumentují tím, že i náhodné texty vykazují Zipfovu distribuci slov. To je však „pouhé přiblížení se skutečnému histogramu ranku. Nejlepší kandidát na skutečnou distribuci ranku zůstává nadále nezodpovězenou otázkou ze dvou důvodů: za prvé test dobré shody zajištěn původním vzorcem je ve statistickém smyslu diskutabilní."[156] S tím se Rerrer-i- Cancho s kolegyní Elvevåg ve své práci vypořádávají tak, že vyhodnocují test dobré shody náhodných textů do skutečných textů přímo prostřednictvím vzorků ranků vytvořených reálných procesem.[157]

„Pokud víme, v žádném z těch populárních článků, které argumentují proti smysluplnosti Zipfova zákona není dostatečně přesná derivace Zipfova zákona z náhodných textů. Tohle má rozhodující význam, protože skutečné texty a náhodné texty se mohou zdát, že mají shodné distribuce ranků, pokud na ně není pohlíženo s dostatečnou precizností jednoduše proto, že dva odlišné drobné objekty můžou vypadat podobně, pokud náš objektiv není dostatečně výkonný." [158] V konstrastu s předchozím článkem Wentiana Li zaznívá, že původní Zipfův zákon je zákonem individuálních ranků, nikoli ranků vybraných k reprezentaci všech slov o stejné délce. Také není jasné, zdali distribuce ranků v textu nesouvisí úzce také s délkou analyzovaného textu, či zda distribuce získaná v kontextu velmi dlouhého textu by byla táž jako distribuce získaná z totožně dlouhého náhodného textu. Z toho důvodu v tomto článku porovnávají autoři skutečné a náhodné texty stejné délky. „Takto můžeme zařídit, že domnělé rozdíly nemohou být připsány jednoduše jen rozdílům v délce textů."[159]

Zaznívá také nespokojenost s nedostatečnou vizuální komparací mezi histogramy skutečných a náhodných textů a proto autoři do svého článku zařadili i část, která se tomuto porovnání zrakem věnuje. Jejich porovnání bude podle všeho odlišné od ostatních autorů, neboť v tomto případě do nej nebudou zahrnuta prázdná slova,

[156] FERRER-i-CANCHO, Ramon, ELVEVAG, Brita. Random texts do not exhibit the real Zipf's law-like rank distribution. In *PLoS ONE 5*, 2010, str. 2.
[157] Tamtéž.
[158] Tamtéž.
[159] Tamtéž.

které některé systémy vytváření náhodných textů také generují (Wentian Li mimochodem toto ve svém způsobu generování náhodného textu nedovoloval). „Konkrétně vizuálně porovnáváme histogramy ranků z anglických textů s těmi z odlišných verzí modelu náhodného textu a pečlivě testujeme test dobré shody náhodných textů na skutečném histogramu z množiny deseti textů. Názorně předvedeme, že - na rozdíl od toho, co bylo dříve naznačováno - náhodné texty selhávají v tom, aby těm skutečných odpovídali alespoň vizuálně.“[160]

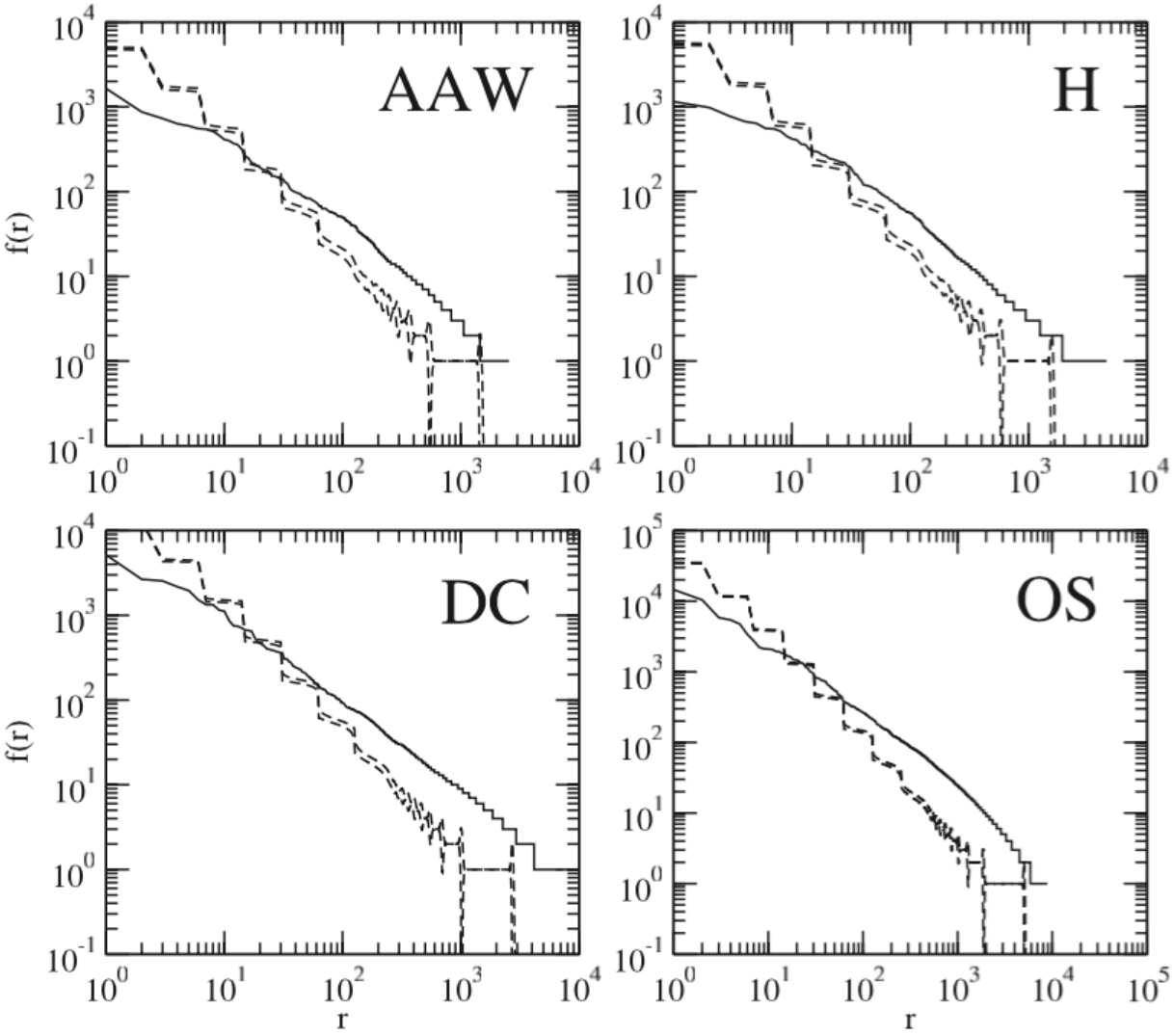
Celá analýza vědců Ferrer-i-Cancho a Elvevåg tedy ve své podstatě stojí na porovnání výsledků získaných touž analýzou, ale jednou na náhodném textu, jednou na textu psaném přirozeným jazykem. Své výsledky podporují jak zmíněnou vizuální komparací, tak i pečlivou aplikací statistických metod. Ale protože všichni autoři textů zařazených do této práce demonstrovali své výsledky primárně vizuální formou grafického zpracování, bude lépe na tuto cestu navázat, díky čemuž se zachová určitá kontinuta. Proto v této práci nebude věnován prostor statistické metodě a jejím výsledkům. Autoři nejprve museli sestavit množinu anglických textů - osm románů, dva eseje [161] - a vybrat si způsoby, kterým si vygenerují texty pro další analýzu. Těmi se stala varianta náhodného textu RT_1, ve které platí, že všechny znaky, prázdná místa nevyjímaje, jsou stejně pravděpodobné. V druhé variantě RT_2 jsou také všechny znaky stejně pravděpodné, ale už jsou z této základní podmínky odebrána právě prázdná místa, aby se zabránilo vznikání prázdných slov, kdy se za sebou mohlo objevit více prázdných míst najednou.

Autoři ve všech variantách dospějí k závěru, že distribuce ranku u náhodných textů a u autentických textů si nijak výrazně neodpovídají. Jako příklad zařazuji výsledky analýzy porovnávající anglické texty a náhodně vygenerované texty metodou RT_1. „Teoretická distribuce ranku nebo dokonce teoreticky očekávaný histogram ranku z náhodných textů není k dispozici, dokonce ani svých nejjednodušších verzích. Z toho důvodu pracujeme na očekávaném histogramu ranku z náhodných textů, který může být snadno odhadnut simulací procesu a průměrováním histogramu ranku přes dostatečně

[160] Tamtéž.

[161] Konkrétně autoři vybrali tato díla: *Alice's adventures in wonderland* (Lewis Caroll), *The adventures of Tom Sawyer* (Mark Twain), *A Christmas carol* (Charles Dickens), *David Crockett* (John S. C. Abbott), *An enquiry concerning human understanding* (David Hume), *Hamlet* (William Shakespeare), *The Hound of the Baskervilles* (Sir Arthur Conan Doyle), *Moby Dick: or, the whale* (Herman Melville), *The origin of species by means of natural selection* (Charles Darwin), *Ulysses* (James Joyce)

velký počet uměle vytvořených textů.“[162] K porovnání si autoři ze své množiny anglických textů vybrali čtyři knihy – *Alice's adventures in wonderland* (AAW), *Hamlet* (H) , *David Crockett* (DC), *The origin of species by means of natural selection* (OS). K vizuální komparaci pak měli autoři tyto výsledky:[163]



[162] FERRER-i-CANCHO, Ramon, ELVEVAG, Brita. Random texts do not exhibit the real Zipf's law-like rank distribution. In *PLoS ONE 5*, 2010, str. 5.
[163] Tamtéž, str. 3.

„Z vizuální inspekce je evidentní, že shoda mezi náhodným textem a reálným textem je chabá. Histogram náhodných textů je zřetelně nad jemu odpovídajícími skutečnými histogramy pro malé ranky a očividně je pod většími ranky.“[164]

Žádná z dílčích analýz v textu Ferrer-i-Cancha a Elvevåg nepřinesla výraznější shodu mezi výsledky reálných textů a textů stvořených zcela nahodile, ať už podle jakéhokoli klíče. A protože jako reálné texty použili díla britské literatury, „za jedno z vysvětlení považují i osobnosti samotných spisovatelů, protože ti nikdy nevytvářejí slova zřetězováním nezávislých událostí za určité pravděpodobnosti. Skuteční spisovatel získává slova z mentálního slovníku, který poskytuje slova ‚připravená k užití‘.“[165] A snad i z tohoto důvodu, který pramení ze samotné podstaty způsobu, jakým náš mozek pracuje s jazykem, si autoři netroufají prohlásit svá zjištění za konečná a obecně platná. Zkorigovali to, co u kolegů považovali za chyby nedovolující nalézt relevantní odpověď, ale sami svou práci uzavírají slovy, že dva fundamentální výzkumné problémy týkající se Zipfových zákonů - tedy jeho smysluplnost a její realistické vysvětlení - zůstávají otevřené.[166]

[164] FERRER-i-CANCHO, Ramon, ELVEVAG, Brita. Random texts do not exhibit the real Zipf's law-like rank distribution. In *PLoS ONE 5*, 2010, str. 5.
[165] Tamtéž, str. 8.
[166] Tamtéž, str. 9.

Závěr

Tématem této diplomové práce byly Zipfovy zákony, základní matematicko-lingvistické principy, které daly vzniknout oboru matematická lingvistika. Cílem práce bylo vytvořit náhled na některé směry bádání v základních jazykových jednotkách jakými jsou písmena, slova, ustálená slovní spojení, texty, korpusy textů a samotná komunikace. Práce je komparativní kompilací vědeckých textů publikovaných v anglickém jazyce.

Nejprve byl představen autor Zipfových zákonů - George Kingsley Zipf, jehož životopis byl čerpán z článku, ke kterému přispěl jeho vlastní syn. Byly nastíněny osobnosti vědců, kteří svým vědeckým zaměřením nebyli jen lingvisté, ale ve své práci v určitém bodě navázali či zareagovali na Zipfovy zákony. Následující oddíl byl věnován zákonům v kontextu vět, korpusů a celé lidské komunikace. Například článek Luďka Hřebíčka, který byl zařazen do této kapitoly, se soustředil na problematiku snahy ověřovat platnost Zipfových zákonů na rozsáhlých textech, ve kterých se podle něj zcela vytrácí lexikální podstata jazyka a proto samotnou velikost vzorku nepovažuje za důležité či šťastné kritérium. Naopak Ramon Ferrer-i-Cancho a Brita Elvevåg ve svém článku týkajícím se náhodných textů vyslovili domněnku, že pro nejspolehlivější porovnání Zipfových zákonů v přirozeném jazyce a v náhodných textech je rozsah přirozených textů zařazených do analýzy. V kapitole popisující distribuční chování slov a slovních spojení byl dán prostor dvěma článkům s protikladnými názory na to, zda Zipfovy zákony platí pro slova nebo pro fráze. Jake Ryland Williams došel se svými kolegy k závěru, že zákony platí pouze pro fráze, Le Quan Ha s kolegy uzavřel svou analýzu výsledkem, že se podle Zipfových zákonů "chovají" slova i fráze.

V poslední kapitole zaměřené na náhodné texty došlo k obdobné situaci - podle Wentiana Li náhodně generované texty vykazují vzorce chování dle Zipfových zákonů, Ramon Ferrer-i-Cancho a Brita Elvevåg se to na základě svých výpočtů nedomnívají. Ačkoli se oba dva texty věnovaly anglickému jazykovému prostředí, do práce byl zařazen i seznam metod pro vygenerování náhodného textu v kontextu českého jazyka, který je i v této oblasti specifický. Spolu s nimi byly popsány i čtyři veřejně dostupné systémy na vytváření náhodných textů, které tyto metody úspěšně užívají v praxi.

Diplomová práce v žádné své části nepřednesla konečnou odpověď na kteroukoli z položených otázek a ani o to neusilovala. Cílem nebylo zvolit „správné

řešení" ani závěry představených článků vyvracet. Cílem diplomové práce bylo předestřít čtenáři z českého prostředí směry, kterými se na poli těchto pár témat ubírají světoví vědečtí pracovníci. Záměrem bylo představit nejen jejich závěry, ale i jejich argumentaci, aby čtenář mohl sám posoudit, nakolik je pro něj který výsledek relevantní. Tyto cíle diplomová práce naplnila, rozhodně však téma neuzavřela, jelikož tohoto závěru zatím nedosáhli ani samotní vědci.

Literatura a zdroje

Použitá literatura

COROMINAS-MURTA, Bernat, Jordi FORTUNY ANDREU a Ricard Vincente SOLÉ. Emergence of Zipf's law in the evolution of communication. *Physical review*. 2011, (83).

ČECH, Radek, Ioan-Iovitz POPESCU a Gabriel ALTMANN. *Metody kvantitativní analýzy (nejen) básnických textů*. 2014.

DAROONEH, A. H. a B. RAHMANI. Finite size correction for fixed word length Zipf analysis. *The European physical journal*. 2009.

HA, Le Quan, Elvira SICILIA-GARCIA, Ji MING a Jack SMITH. *Extension of Zipf's Law to Words and Phrases*. 2002.

HŘEBÍČEK, Luděk. Zipf's law and text. *Glottometrics*. 2002, (3).

FERRER-I-CANCHO, Ramon a Brita ELVEVÅG. Random texts do not exhibit the real Zipf's law-like rank distribution. *PLoS ONE*. 2010, (5).

LI, Wentian. Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on information theory*. 1991, (38).

LYER, Stanislav. Slovo a jeho struktura. *Slovo a slovesnost*. 1942, **8**(1).

MANDELBROT, Benoît. An informational theory of the statistical structure of language. *JACKSON, Willis. Communication theory: Papers read at a symposium on "Applications of communication theory" held at the Institution of electrical engineers, London September 22nd-26th 1952*. 1953.

PRUN, Claudia a Robert ZIPF. Biographical notes on G. K. Zipf. *Glottometrics*. 2002, (3).

ROUSSEAU, Ronald. George Kingsley Zipf: life, ideas, his law and informetrics. *Glottometrics*. 2002, (3).

SIMON, Herbert. *On a class of skew distribution functions*. 1995.

TĚŠITELOVÁ, Marie. K statistickému výzkumu slovní zásoby. *Slovo a slovesnost*. 1961, **22**(3).

WILLIAMS, Jake Ryland, Paul LESSARD, Suma DESU, Eric CLARK, James BAGROW, Christopher DANFORTH a Peter Sheridan DODDS. Zipf's law holds for phrases, not words. *In Scientific reports*. 2015, (5).

WYLLYS, Ronald. Empirical and theoretical bases of Zipf's law. *In Library Trends*. 2015, **30**(1 Summer).

ZANETTE, Damián H. a Marcelo A. MONTEMURRO. Dynamics of text generation with realistic Zipf distribution. *Journal of quantitative linguistics*. 2005, (12).

ZIPF, George K. *Human behavior and the principle of least effort*. 1949.

ŽOUŽELKOVÁ, Jana. *Nové míry volatility ekonomických časových řad*. Olomouc, 2012. Diplomová práce. Univerzita Palackého v Olomouci. Vedoucí práce RNDr. Tomáš Fürst, Ph.D.

Zmíněná literatura

DITTMAR, Jeremiah. *Cities, Institutions, and Growth: The emergence of Zipf's law*. 2010.

GABAIX, Xavier. Zipf's law for cities: an explanation. *The Quarterly journal of economics*. 1991.

JIANG, Bin a Tao JIA. Zipf's law for all the natural cities in the United States: A geospatial perspective. *In International Journal of Geographical Information Science*. 2010.

MOURA, Newton a Marcelo RIBERIO. Zipf law for Brazilian cities. *Physica A-statistical mechanics and its applications*. 2006.

TĚŠITELOVÁ, Marie. *Otázky frekvence slov (zvláště v češtině),* nepublikovaná disertační práce, 1951, 10-22.

Online zdroje

České lorem ipsum. Blábot. [online]. 2011 [cit. 2016-08-10]. Dostupné z: http://cs.blabot.net/s/d1p1o0u0b20-20s20-20#pro-developery

Chci generovat. lipsum.cz: generátor náhodného textu. [online]. [cit. 2016-08-10]. Dostupné z: http://lipsum.cz/

Generátor náhodného výplňového textu (Dummy Text Generator). Drivel: Dummy text generator. [online]. 2007-2012 [cit. 2016-08-10]. Dostupné z: http://drivel.ikit.cz/

Lorem Ipsum. Co je Lorem Ipsum?. [online]. [cit. 2016-08-10]. Dostupné z: http://cs.lipsum.com/

Makra v MS Wordu, čárový kód. Moderní techniky programování 1: Cvičení č. 6. [online]. 13.6.2006 [cit. 2016-08-10]. Dostupné z: http://www.fce.vutbr.cz/aiu/vojkuvka.m/6u4/cviceni06.htm

Teorie velkého blábolu. Blábot. [online]. 2011 [cit. 2016-08-10]. Dostupné z: http://cs.blabot.net/s/d1p1o0u0b20-20s20-20#pro-developery

Tvorba univerzálních projevů. KÝBLsoft: …Geniální algoritmy pro každého. [online]. 1995-2016 [cit. 2016-08-10]. Dostupné z: http://www.kyblsoft.cz/projevy

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 1978". Nobelprize.org. Nobel Media AB 2014. Web. 20 Apr 2016. <http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/1978/>

Vygenerovaný projev. KÝBLsoft: …Geniální algoritmy pro každého. [online]. 1995-2016 [cit. 2016-08-10]. Dostupné z: http://www.kyblsoft.cz/projevy?zacatek_projevu=Soudruzi&pocet_vet_o_hovne=20&pocet_odstavcu=3